PAPER

Analysis of seismic inversion with optimal transportation and softplus encoding

IOP Publishing Inverse Problems
Inverse Problems **37** (2021) 095004 (24pp) https://doi.org/10.1088/1361-6420/ac1511

Analysis of seismic inversion with optimal transportation and softplus encoding

Lingyun Qiu *

Yau Mathematical Sciences Center, Tsinghua University, Beijing, People's Republic of China
Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, People's Republic of China

E-mail: lyqiu@tsinghua.edu.cn

Received 15 December 2020, revised 22 June 2021
Accepted for publication 9 July 2021
Published 4 August 2021

Abstract

This paper is devoted to the theoretical and numerical investigation of the local minimum problem in an inverse boundary value problem. We provide a mathematical analysis for the objective function of the optimal transportation type in the context of seismic full waveform inversion. In particular, we prove that the gradient obtained using the adjoint-state method does not depend on the specific choice of the Kantorovich potentials. Moreover, our frequency analysis results show a decreasing sensitivity of the reconstruction as the data misfit is concentrated in the high-frequency part. This confirms previous observations in many numerical experiments. We also propose a new method using the softplus encoding, which maps the seismic data into probability measures and, therefore, the Wasserstein metric can be applied. The softplus encoding retains the convexity of the data misfit with respect to translations and provides a parameter to tune the landscape of the objective function. The effectiveness of the proposed method is demonstrated numerically on an inversion task with the benchmark Marmousi model.

Keywords: inverse problems, full waveform inversion, optimal transportation, quadratic Wasserstein distance

(Some figures may appear in colour only in the online journal)

*Author to whom any correspondence should be addressed.

1. Introduction

Seismic full waveform inversion (FWI) uses kinematic and dynamical information of the seismic wavefield to build the subsurface velocity model, which accurately depicts the geological structures. Mathematically, FWI is formulated as a nonlinear inverse problem matching modeled data to the recorded field data [1]. It can be solved as a PDE-constrained optimization problem, and a least-squares objective function is used for measuring the data misfit. The objective function is minimized with respect to the model parameter, and the model update is computed using the adjoint state method [2]. FWI can produce high-resolution models of the subsurface compared to ray-based methods. However, FWI is often an ill-posed problem due to the band-limited nature of the seismic data and the limitations of the acquisition geometries.

The least-squares formulation of FWI, when the initial model is far from the true model and the seismic data lack of low-frequency information, tends to produce many local minima. That is the so-called cycle-skipping issue. The cause of this issue is that only the pointwise amplitude difference is measured with the L_2 norm, while the phase or travel-time information embedded in the data is more critical for the inversion. Different approaches are proposed to capture more accurate kinematic information, such as dynamic time warping and convolution-based methods. This information is then used to optimize the objective function or enlarge the true solution valley. In this direction, we mention the works in [3–7]. An alternative approach to reshape the objective function is to extend the parameter space [8, 9] or use an auxiliary wavefield [10–12] in a non-physical way so that the data can be easily fitted. Then, one can get the physical model back using an annihilator or gradually tightening the PDE constraints.

Another approach involves the use of Wasserstein metrics. The Wasserstein distance and optimal transportation (OT) theory were first brought up to seek the optimal cost of rearranging one density into the other, where the transportation cost per unit mass is the Euclidean distance or Manhattan distance. It can be traced back to the mass transport problem proposed by Monge in 1780s and its relaxed formulation by Kantorovich in the 1940s. Since then, it has become a classical subject in probability theory, economics, computer vision, optimization, and partial differential equations.

Recently, the Wasserstein distance and its variants are proposed to replace the distance for the objective function in FWI. The idea of using OT metrics for seismic inverse problems is introduced in [13]. The successful applications are partly due to its Lagrangian nature that focuses on the trajectory of individual masses. This makes the Wasserstein metric a vital tool in capturing the variations of signals as a whole, such as translation (time-shift) and dilation. However, Wasserstein metrics cannot be directly applied to the seismic inversion due to the oscillatory and sign-change behavior of the seismic data. To overcome this difficulty, different approaches have been proposed in the literature. The first category is to transfer the data to a probability measure. In [13–19], linear, quadratic, exponential transforms, and a map that separates the data into its negative and positive parts, followed by normalization, are used to turn data into a probability measure. These encoding methods have their drawbacks. For example, the linear and exponential transforms introduce new masses everywhere and therefore destroy the convexity with respect to large translation. The negative/positive separation method retains the convexity, but it is sensitive to noise, and it creates numerical dispersion significantly. As for the quadratic transform, there is still no report of its successful application in inverse medium problems as far as we know. In [20], a graph-based transform is introduced, which maps the data $u(t)$ to a normalized Dirac measure, $c\delta(\{t, u(t)\})$, and the OT metric is calculated on the

graph space. This approach extends the dimensionality of the space by one and cannot take advantage of the closed-form solution of one-dimensional OT.

Another category is to relax the non-positiveness and equal-mass restrictions of the Kantorovich–Rubinstein distance, the Wasserstein metric with cost function $|x-y|$. In [20–22], the authors use a Kantorovich–Rubinstein norm defined on the space of Borel measures [23, 24], which restricts the supremum of φ's in the Kantorovich–Rubinstein duality,

$$W_1(\mu,\nu) = \sup_{\|\varphi\|_{\mathrm{Lip}}\leqslant 1} \int \varphi \mathrm{d}(\mu-\nu)$$

to functions satisfying $0 \leqslant \varphi \leqslant M$ for some given constant M. To obtain a computationally feasible solver for the multi-dimensional case, a simultaneous-direction method of multipliers is proposed [21]. The main drawback of the Kantorovich–Rubinstein approach is the loss of convexity with respect to large translations [20]. In this direction, we also mention the works using $\dot{\mathcal{H}}^{-1}$ [25], unbalanced OT distances [26] and matching filters [27].

Among various strategies for mitigating the cycle-skipping issue in seismic inversion, using OT-based objective functions has been demonstrated to be one of the most effective approaches. However, these approaches have three points that require further investigation. First of all, there is sufficient evidence that FWI has a strong path dependence. However, the Kantorovich potential generally does not have uniqueness. Hence the associated model gradient obtained using the adjoint-state method is not unique. Secondly, the inversion results depend crucially on the appropriate underlying encoding method to transfer the seismogram to probability density functions (PDFs). Last but not least, OT metrics mainly concern the cycle-skipping problem. It thus tends to give smooth inversion results and lacks high-resolution details. In order to delineate usage scenarios, frequency sensitivity analysis is crucial.

Our goal of the present paper is to provide a rigorous description of the gradient-based methods and proper encoding methods for the seismic inverse problem using OT. These findings are essential for a rigorous interpretation of the numerical observations. Several objectives are discussed in this paper. First, a rigorous proof is presented on the directional differentiability of the transportation cost as a function in L_2. We also perform a frequency sensitivity analysis of the OT objective function using the Fourier series. Second, an encoding method using the softplus function is introduced, and then it is proved that the gradient obtained using the adjoint state method is well-defined and unique. Finally, by applying it to a simple convexity test and an inverse problem on the benchmark Marmousi model, the feasibility of the proposed method is demonstrated.

The paper is organized as follows. The necessary notations and properties of the quadratic Wasserstein distance, especially the efficient solution in the unidimensional case, are discussed in section 2. As shown in section 3, one only needs to change the adjoint source when switching from L_2 to other metrics in the objective function, and an adjoint source involving the first Kantorovich potential and the gradient of the encoding map is used for the OT one. In section 4, we investigate the frequency sensitivity of OT and specify a low-frequency enhancement of it; a rigorous proof of the directional differentiability and uniqueness of the gradient is also presented. The desired properties of encoding methods and an effective approach using the softplus function are illustrated in section 5. Two numerical examples are shown in section 6. For completeness and reproducibility of the results, the pseudo-code of the proposed method is presented in appendix A.

2. The quadratic Wasserstein distance

This section introduces the quadratic Wasserstein distance used to measure the difference between data. We begin with some standard notations and necessary properties.

2.1. Notation

Throughout this paper, we shall consider probability measures that are absolutely continuous with respect to the Lebesgue measure and with a finite moment of order 2 on a simply connected and compact domain in $\mathbb{R}^n$. Hence, we identify the induced measure with its Radon–Nikodym derivative with respect to the Lebesgue measure and write $\mathrm{d}\mu(x) = \mu(x)\mathrm{d}x$. The measure and its Radon–Nikodym derivative will not be distinguished, as it should be clear from the context. All the measures considered here are built from the solution to wave equations. The regularity condition is clearly satisfied, and the limited-time/space measurement of the data leads to the boundedness of the domain. When no ambiguity arises, we denote for brevity by $\mathcal{P}$ the set of all absolutely continuous measures with a finite moment of order 2 on the given domain.

In this work, we use Kantorovich's formulation of OT and its dual form. The definition and properties are summarized in the following. We refer readers to [23, 28, 29] for a more detailed discussion.

Definition 2.1 (Kantorovich's OT problem). Let $\mu, \nu \in \mathcal{P}$. Minimize

$$I[\gamma] = \int \frac{1}{2}|x-y|^2 \,\mathrm{d}\gamma(x,y) \tag{1}$$

over the set of all coupling measures, which admit μ and ν as marginals on the first and second factors respectively, i.e.

$$\int (\varphi(x)+\psi(y))\,\mathrm{d}\gamma(x,y) = \int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu, \tag{2}$$

for all measurable functions $\varphi \in L^1(\mathrm{d}\mu)$ and $\psi \in L^1(\mathrm{d}\nu)$.

Theorem 2.1 (Kantorovich duality). *The minimum of Kantorovich's problem* (1) *is equal to the supremum of*

$$\int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu \tag{3}$$

over all pairs $(\varphi,\psi) \in L^1(\mathrm{d}\mu)\times L^1(\mathrm{d}\nu)$ *such that* $\varphi(x)+\psi(y) \leqslant \frac{1}{2}|x-y|^2$.

The supremum in theorem 2.1 is attainable and the Wasserstein distance between μ and ν is defined as

$$W_2(\mu,\nu) = \min_{\gamma} I[\gamma]^{1/2}. \tag{4}$$

For simplicity, we consider the second power of W_2, which is the optimal total transportation cost $\mathcal{T}(\mu,\nu) = W_2^2(\mu,\nu)$.

2.2. OT on the real line

The one-dimensional case is of particular interest, as its equivalent definition does not involve solving a minimization problem. It can be solved explicitly and efficiently with a linear

computational complexity from a computational point of view. From a theoretical point of view, the 1D Wasserstein distance is strongly convex along the geodesic as a function of its first argument. All high-dimensional ones are not even convex along the geodesic (see, e.g. [30, example 9.1.5]). From the perspective of the seismic inverse problem, this leads us to consider using the OT metric on the time variable combined with the least-squares on the spatial variable, rather than the high-dimensional Wasserstein distances on the space-time variable.

The following theorem states a solution to the Monge–Kantorovich problem on the real line in terms of cumulative distribution functions.

Theorem 2.2 (OT theorem on $\mathbb{R}$ [23, 29]). *Let $p_0, p_1 \in \mathcal{P}(\mathbb{R})$ be two probability measures on the real line. f_0 and f_1 are their cumulative distribution functions:*

$$f_k(x) = \int_{-\infty}^{x} \mathrm{d}p_k, \quad k = 0, 1.$$

The pseudo-inverse of a non-decreasing and right-continuous function f is defined by

$$f^{[-1]}(x) = \inf\{t \in \mathbb{R} \mid f(t) > x\}.$$

Then, there exists a unique non-decreasing map $T : \mathbb{R} \to \mathbb{R}$ given by $T(x) = f_0^{[-1]}(f_1(x))$ such that $p_0(T(x)) = p_1(x)$. The map T is optimal in the Monge–Kantorovich problem for the quadratic cost function. Moreover, the value of the optimal transport cost is

$$\begin{aligned} W_2(p_0, p_1) &= \left(\int_0^1 (f_0^{[-1]}(s) - f_1^{[-1]}(s))^2 \mathrm{d}s \right)^{1/2} \\ &= \left(\int_{-\infty}^{+\infty} (f_0^{[-1]}(f_1(t)) - t)^2 p_1(t) \mathrm{d}t \right)^{1/2}. \end{aligned} \tag{5}$$

Remark 2.3. There are several aspects to be mentioned here regarding the optimal transport map, T. First, if the two measures are atomless and strictly positive, and hence, the cumulative distribution functions are continuous and strictly monotone, then one would have

$$T = f_0^{-1} \circ f_1.$$

Second, from the explicit form of T, we conclude that the regularity of T is one degree higher than that of the measures. Higher regularity leads to a smoother effect. This has been observed in the numerical experiments. Third, the form of T implies that a monotone rearrangement of p_0 gives the solution to the transportation problem onto p_1. This leads to the algorithm with computational cost $O(N)$ for computing the transportation cost and its first variation. Please refer to appendix A for more details. Fourth, the transportation map T is optimal not only for the quadratic cost, but also for all cost functions in the form of $c(x, y) = h(y - x)$ with h being a convex function. In particular, the OT cost associated with the cost function $c(x, y) = |x - y|$ is

$$W_1(p_0, p_1) = \int_0^1 \left| f_0^{[-1]}(s) - f_1^{[-1]}(s) \right| \mathrm{d}s = \int_{-\infty}^{+\infty} |f_0(x) - f_1(x)| \, \mathrm{d}x.$$

Finally, the first variation of the transportation cost is given by

$$\frac{\partial W_2^2(p_0,p_1)}{\partial p_0} = (f_1^{[-1]}(f_0(t)) - t)^2 + \int_t^1 2\frac{\partial f_1^{[-1]}(x)}{\partial x}\bigg|_{x=f_0(s)} \times \left(f_1^{[-1]}(f_0(s)) - s\right) p_0(s)\mathrm{d}s. \tag{6}$$

To simplify the calculation and avoid the differentiation, which may cause some numerical error, the second term in the above formula can be rewritten as

$$2\int_{f_1^{[-1]}(f_0(t))}^1 \left(s - f_0^{[-1]}(f_1(s))\right)\mathrm{d}s. \tag{7}$$

Here the inverse function theorem is applied.

3. Full waveform inversion

In this section, we first briefly review the theory of FWI and the adjoint state method. Then, an analog of the adjoint wavefield using transportation distance is developed. The differentiability and uniqueness will be analyzed in subsequent sections. In section 6, it will be used in conjunction with the softplus encoding method to perform numerical experiments.

We start with the acoustic wave equation in the time domain governed by

$$\left(m(x)\frac{\partial^2}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho(x)}\nabla\right)\right)u(x,t) = f(x,t), \tag{8}$$

where m is the reciprocal of the bulk modulus, ρ is the density, and u and f stand for the pressure wavefield and source term, respectively. We symbolize the relationship between the model parameters and the observed wavefield by an operator F, which is also referred to as the forward operator,

$$F(m,\rho,f) = u|_\Gamma. \tag{9}$$

Γ stands for the receiver geometry, which is usually a portion of a surface or a collection of discrete points.

The goal of the inverse problem is to reconstruct the model parameters from the measured data. Usually, the inverse problem is posed as a nonlinear least-squares optimization problem,

$$\min_{m,\rho,f}\mathcal{J}(m) = \frac{1}{2}\|F(m,\rho,f) - \mathrm{d}\|_2^2, \tag{10}$$

where $\mathcal{J}$ is the misfit function and $\|\cdot\|_2$ is the L_2 norm. That is, to choose the model parameters such that the correspondingly simulated waveform yields the minimum difference away from the measured data in the L_2 sense. For simplicity, we assume the density ρ and source term f are known in this work. Hence we omit the explicit dependence on ρ, f in (9) and (10) in the following sections.

3.1. Adjoint state method

Modern techniques for seismic inverse problems involve using data with sizes ranging from gigabytes to terabytes or even petabytes. The adjoint state method plays a significant role in

the computational aspect of large-scale optimization problems. For completeness, a simple description is included here in a general setting. For more details on this topic, please refer to [2].

Suppose the misfit function is $\mathcal{J}(u(m))$, where u and m stand for the state variable and model parameter, respectively. u and m satisfy the state equation $\Phi(m, u(m)) = 0$. For the gradient-based method, the total derivative $\delta\mathcal{J}/\delta m$ needs to be computed to assess the sensitivity of the misfit function to the model parameter. The gradient $\delta\mathcal{J}/\delta m$ is simply

$$\frac{\delta\mathcal{J}}{\delta m} = \left\langle \frac{\delta\mathcal{J}}{\delta u}, \frac{\delta u}{\delta m} \right\rangle, \tag{11}$$

where the inner product acts in the space of u, and $\delta u/\delta m$ is a linear operator acting on perturbations on m and returning perturbations on u. In the context of FWI, the difficulty of numerically evaluating $\delta\mathcal{J}/\delta m$ lies in the evaluation of the wavefield perturbation δu for all possible model perturbation δm. The adjoint state method answers the question, 'how to efficiently calculate $\delta\mathcal{J}/\delta m$ without evaluating $\delta u/\delta m$ explicitly?'

Let us define the adjoint state variable v as the solution of the adjoint state equation,

$$\left(\frac{\partial\Phi}{\partial u}\right)^{*} v = \frac{\delta\mathcal{J}}{\delta u}. \tag{12}$$

From the state equation, we know that

$$\frac{\partial\Phi}{\partial u}\frac{\delta u}{\delta m} + \frac{\partial\Phi}{\partial m} = 0. \tag{13}$$

It follows that

$$\begin{aligned}
\frac{\delta\mathcal{J}}{\delta m} &= \left\langle \frac{\delta\mathcal{J}}{\delta u}, \frac{\delta u}{\delta m} \right\rangle \\
&= \left\langle \left(\frac{\partial\Phi}{\partial u}\right)^{*} v, \frac{\delta u}{\delta m} \right\rangle \\
&= \left\langle v, \frac{\partial\Phi}{\partial u}\frac{\delta u}{\delta m} \right\rangle \\
&= \left\langle v, -\frac{\partial\Phi}{\partial m} \right\rangle.
\end{aligned} \tag{14}$$

In the above identities, we omit the explicit dependence of the inner products on the associated spaces for simplicity. Indeed, from this formulation, one observes that as long as the wavefield u is still the intermediate in the construction of the data misfit function $\mathcal{J}$, only the adjoint state variable v depends on the specific form of $\mathcal{J}$. Furthermore, only the adjoint source term $\frac{\delta\mathcal{J}}{\delta u}$ needs to be modified for different misfit functions as long as it is of the form $\mathcal{J} = \mathcal{J}(u(m))$.

In the conventional FWI with least-squares misfit function

$$\mathcal{J} = \frac{1}{2}\|u - \mathrm{d}\|_2^2, \quad u = F(m),$$

we have that

$$\frac{\delta\mathcal{J}}{\delta u} = u - \mathrm{d}.$$

Applying the adjoint state method gives

$$\frac{\delta \mathcal{J}}{\delta m} = \left\langle v, -\frac{\partial \Phi}{\partial m} \right\rangle, \tag{15}$$

where the adjoint state variable v solves the adjoint state equation

$$\begin{cases} \left(m(x)\dfrac{\partial_2}{\partial t^2} - \nabla \cdot \left(\dfrac{1}{\rho(x)} \nabla \right) \right) v(x,t) &= u - \mathrm{d}, \\ v(x,T) &= 0, \\ \partial_t v(x,T) &= 0. \end{cases} \tag{16}$$

For the FWI with quadratic Wasserstein norm and proper encoding, the data misfit function is defined as

$$\mathcal{J} = W_2^2(\tilde{u}, \tilde{\mathrm{d}}), \quad u = F(m), \tilde{u} = \mathcal{D}(u), \tilde{\mathrm{d}} = \mathcal{D}(\mathrm{d}), \tag{17}$$

where $\mathcal{D}$ is the encoding operation from seismic data to equal-mass non-negative measures. It follows that

$$\frac{\delta \mathcal{J}}{\delta u} = \left\langle \frac{\mathrm{d}W_2^2(\tilde{u}, \tilde{\mathrm{d}})}{\mathrm{d}\tilde{u}}, \frac{\mathrm{d}\mathcal{D}(u)}{\mathrm{d}u} \right\rangle = \mathcal{D}'[u]^*(\varphi), \tag{18}$$

where φ is the Kantorovich potential of $W_2^2(\tilde{u}, \tilde{\mathrm{d}})$ associated with $\tilde{u}$. Then, applying the adjoint state method, we obtain that

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}m} = \left\langle v, -\frac{\partial \Phi}{\partial m} \right\rangle, \tag{19}$$

where the adjoint state variable v solves the adjoint state equation

$$\begin{cases} \left(m(x)\dfrac{\partial^2}{\partial t^2} - \nabla \cdot \left(\dfrac{1}{\rho(x)} \nabla \right) \right) v(x,t) &= \mathcal{D}'[u]^*(\varphi), \\ v(x,T) &= 0, \\ \partial_t v(x,T) &= 0. \end{cases} \tag{20}$$

4. Wasserstein metric from a seismic inverse problem perspective

In this section, we discuss the features of the quadratic Wasserstein metric from a seismic inverse problem perspective. We start by investigating the frequency sensitivity of $\mathcal{T}$. It is proved that $\mathcal{T}$ emphasizes the low-frequency components not only locally in the linearization regime but also in a global sense. This also reveals that the sensitivity of the solution is minor in highly oscillating data. Next, we present the rigorous definition of a set, says $\mathcal{D}$, in which the optimization is performed. We show the Euclidean differentiability of the transportation cost and that the gradient is unique up to an additive constant for any element in $\mathcal{D}$. This set will be used as a desirable image domain to design the encoding mapping.

4.1. Frequency sensitivity of W_2

A long-standing view in seismic inversion starts with low-frequency data, which contain large-scale, kinematically relevant components of the velocity model. The low-to-high frequency-continuation schemes [31–35] help FWI mitigate the cycle-skipping issue, i.e. the local minimum problem. At the same time, an overly detailed frequency division will slow down the entire inversion process significantly. As is well known, the quadratic Wasserstein distance $W_2(\mu, \cdot)$ is asymptotically equivalent to a weighted $\dot{H}^{-1}(\mathrm{d}\mu)$, where $\dot{H}^{-1}$ denotes the dual space of the space of zero-mean H^1 function. It is also well known that L_2 measures different frequency components equally, and $\dot{H}^{-1}$ attenuates them with a polynomial weight of order $|k|^{-1}$. The following theorem shows a non-asymptotically similar behavior of W_2 and $\dot{H}^{-1}$.

Theorem 4.1. *Assume that $\mu_0, \mu_1 \in \mathcal{P}(S^1)$, where S^1 stands for the unit circle, and*

$$\mu_1 = \mu_0 + \sum_{k\in\mathbb{Z}^+} \left(a_k \cos(k\theta) + b_k \sin(k\theta)\right). \tag{21}$$

Note that the 0-frequency amplitude vanishes since $\int \mathrm{d}\mu_0 = \int \mathrm{d}\mu_1$. If

$$\nu = \mu_0 - \sum_{k\in\mathbb{Z}^+} \left((a_k \cos)^-(k\theta) + (b_k \sin)^-(k\theta)\right), \tag{22}$$

is a non-negative measure on S^1, then

$$W_2^2(\mu_0, \mu_1) \leqslant \sum_{k\in\mathbb{Z}^+} \frac{2\pi^2}{k^2} \left(|a_k| + |b_k|\right). \tag{23}$$

Here, f^- stands for the negative part of the Radon measure f.

Proof. We shall find at least one (*a priori* not optimal) transport plan from μ_0 to μ_1 by rearranging only $|a_k|$ or $|b_k|$ mass within an arc of length $2\pi/k$. Let

$$D_k = \left\{(\theta, \varphi) \in S^1 \times S^1 | \varphi = \left(\theta + \frac{\pi}{k}\right) \bmod 2\pi\right\}, \quad k \in \mathbb{Z}^+,$$

and D_∞ be the diagonal $\{(\theta, \theta)\}$ in $S^1 \times S^1$. Consider the following coupling:

$$\kappa = \delta(D_\infty)\nu + \sum_{k\in\mathbb{Z}^+} \delta(\mathrm{d}_k) \left((a_k \cos)^+(k\theta) + (b_k \sin)^+(k\theta)\right). \tag{24}$$

This coupling keeps an amount of mass in place, which is shared between μ_0 and μ_1, and transport the rest within one corresponding period. It follows that κ has marginals μ_0 and μ_1 and is an admissible transport plan. This means that

$$W_2^2(\mu_0, \mu_1) \leqslant \int_{S^1\times S^1} c(\theta, \varphi)\mathrm{d}\kappa(\theta, \varphi) = \frac{2\pi^2}{k^2} \left(|a_k| + |b_k|\right), \tag{25}$$

where the cost function $c(\theta, \varphi) = \min(|\theta - \varphi|^2, (2\pi - |\theta - \varphi|)^2)$ is associated with the geodesic distance along the circle. □

Remark 4.2. In the proof of theorem 4.1, we use a constructive approach rather than the explicit solution of the 1D OT. The result holds true for high-dimensional domains with boundaries. The proof needs to be modified concerning boundary treatment, and the corresponding

weight is $|k|^{-2}$. It is also worth mentioning that W_2 is not very sensitive to oscillations and offers a natural weighting that emphasizes low-frequency differences. Therefore, the primary motivation for using W_2 is to solve large-scale errors instead of pursuing high-resolution imaging.

4.2. Gradient of quadratic Wasserstein distance

The seismic inverse problem is that of solving for model functions in a nonlinear system. Considering the large scale of the system, the commonly used approach is to formulate the inverse problem as an optimization problem and solve it with gradient-based methods. A brief discussion of the directional differentiability properties of the quadratic Wasserstein distance along certain directions is presented here. We start by extending $\mathcal{J}_\nu(\mu) = \mathcal{T}(\mu, \nu)$ to a functional on L_2.

Roughly speaking, the optimization is performed using linearization in a vector space, and, instead of the L_2-norm, the total transportation cost is used as the objective function. As a result, this suggests that it is necessary to extend the functional from the probability space to the L_2 space. With a slight abuse of notation, we extend the functional to $\mathcal{T} : L_2 \times L_2 \to [0, +\infty]$ by

$$\mathcal{T}(\mu, \nu) = \begin{cases} W_2^2(\mu, \nu), & \text{if } \mu, \nu \in \mathcal{P}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{26}$$

Next, we introduce a subset $\mathcal{U} \subset \mathcal{P}$, which is, in some sense, served as the 'interior' of $\mathcal{P}$. Then, a short discussion is presented on the differentiability properties of the transportation cost $\mathcal{T}(\mu, \nu)$ over $\mathcal{U}$, see [23, 29] for more detail and more general cases. Discussion in this section paves the way to data encoding and minimization of the misfit between seismic data in the transportation sense.

Let Σ be the Borel σ-algebra on the given bounded domain in $\mathbb{R}^n$ and

$$\mathcal{U} = \left\{ \mu \in \mathcal{P} \mid \exists r > 0 \text{ s.t. } \int \chi_A \, \mathrm{d}\mu \geqslant r \int \chi_A \, \mathrm{d}x, \quad \forall\, A \in \Sigma \right\}. \tag{27}$$

Theorem 4.3. *Let $\mathcal{T} : L_2 \times L_2 \to [0, +\infty]$ be the extended transportation cost. Consider the functional $\gamma \mapsto \mathcal{T}(\gamma, \nu)$ for a fixed measure $\nu \in \mathcal{P}$. If $\mu \in \mathcal{U}$, then*

$$\frac{\partial \mathcal{T}(\gamma, \nu)}{\partial \gamma}(\mu) = \varphi, \tag{28}$$

where φ is the Kantorovich potential associated with μ and is unique up to additive constants.

Proof. For some fixed $\gamma \in \mathcal{U}$, consider the sequence $\{\mu_t \triangleq \mu + t(\gamma - \mu)\}$ converging to μ in the sense of

$$\lim_{t \to 0} \mathcal{T}(\mu_t, \mu) = 0. \tag{29}$$

By the triangle inequality on W_2, one gets

$$\lim_{t \to 0} \mathcal{T}(\mu_t, \nu) - \mathcal{T}(\mu, \nu) = 0, \quad \forall\, \nu \in \mathcal{P}. \tag{30}$$

Let (φ, ψ) be an optimizing pair in the Kantorovich dual formulation, i.e.

$$\mathcal{T}(\mu, \nu) = \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu, \tag{31}$$

and we additionally assume that $\int \varphi(x)\mathrm{d}x = 0$, thus making the unique determination of (φ, ψ). The sub-differentiability of $\mathcal{T}(\cdot, \nu)$ follows from the fact that (φ, ψ) is optimal for $\mathcal{T}(\mu, \nu)$, and is not necessarily optimal for $\mathcal{T}(\mu_t, \nu)$,

$$\begin{aligned}\mathcal{T}(\mu_t, \nu) - \mathcal{T}(\mu, \nu) &\geqslant \left(\int \varphi \,\mathrm{d}\mu_t + \int \psi \,\mathrm{d}\nu\right) - \left(\int \varphi \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu\right) \\ &= t \int \varphi \mathrm{d}(\gamma - \mu). \end{aligned} \tag{32}$$

For the other part of the differentiability, we denote a subsequence realizing the limit superior of $\mathcal{T}(\mu_t, \nu)$ by $\{\mu_{t_k}\}$, i.e.

$$\lim_{k \to +\infty} \mathcal{T}(\mu_{t_k}, \nu) = \limsup_{t \to 0} \mathcal{T}(\mu_t, \nu), \tag{33}$$

and let (φ_k, ψ_k) be an optimizing pair for $\mathcal{T}(\mu_{t_k}, \nu)$. Additionally, we assume $\int \varphi_k(x)\mathrm{d}x = 0$. Thus, the uniqueness of φ_k follows by the $\mathrm{d}\mu_k$-uniqueness of $\nabla\varphi_k$ and the fact that $\mu_k \in \mathcal{U}$ is positive. Then, we conclude from the suboptimality of (φ_k, ψ_k) for $\mathcal{T}(\mu, \nu)$ that

$$\begin{aligned}\mathcal{T}(\mu_{t_k}, \nu) - \mathcal{T}(\mu, \nu) &\leqslant \left(\int \varphi_k \,\mathrm{d}\mu_{t_k} + \int \psi_k \,\mathrm{d}\nu\right) - \left(\int \varphi_k \,\mathrm{d}\mu + \int \psi_k \,\mathrm{d}\nu\right) \\ &= t_k \int \varphi_k \mathrm{d}(\gamma - \mu). \end{aligned} \tag{34}$$

From the stability of the optimal transference mapping and Brenier's theorem [23], we know $\varphi_k \rightharpoonup \varphi$. Hence

$$\lim_{t \to 0} \frac{\mathcal{T}(\mu_t, \nu) - \mathcal{T}(\mu, \nu)}{t} = \int \varphi \mathrm{d}(\gamma - \mu). \tag{35}$$

The uniqueness of φ up to additive constants follows by noting that $\nabla\varphi$ is $\mathrm{d}\mu$-unique and μ is positive everywhere.

□

Remark 4.4 (On the strictly positive range of the encoding mapping). Usually, functions differing only on a measure-null set are not distinguished. In the inverse problem context, one compares two encoded data and does not expect them to be invisible to each other. The definition of $\mathcal{U}$ originates from the idea that any two elements of $\mathcal{U}$ should be absolutely continuous to each other and the observation $\mathcal{U} + \varepsilon\xi \subset \mathcal{U}$ for small ε and bounded mean-zero perturbation ξ. On the other hand, the positiveness of μ is required to ensure that the derivative φ is unique up to additive constants over the whole domain. In the next section, this uniqueness will be used to show the associated gradient for the velocity model is unique, and therefore, the adjoint state method is well defined. Last but not least, the uniform lower bound r in the definition of $\mathcal{U}$ is to give more space for the line search in the optimization.

5. Encoding methods

In this section, we investigate the criterion for selecting a proper encoding method to transfer the non-Wasserstein-measurable seismic data into PDFs. A simple but quite useful strategy using the softplus function is presented, and some useful properties are examined. Our goal is to make the data misfit measurable using the Wasserstein distance and efficiently calculate the

associated gradient. In this perspective, we suggest the following strategies to choose encoding map $\mathcal{D}$:

(a) The range of $\mathcal{D}$ is contained in $\mathcal{U}$;
(b) $\mathcal{D}$ is differentiable and invertible;
(c) $\mathcal{D}$ is a pointwise mapping, i.e. $(\mathcal{D} \circ u)(x) = \mathcal{D}(u(x))$.

The first point guarantees the existence and uniqueness (up to an additive constant) of the first variation of the transportation cost. The second one makes the mapping compatible with quasi-Newton type methods. The third point is purely for the sake of efficiency. Usually, to match the mass of the encoded data, a normalization procedure is involved, and it is hard to ensure the invertibility of the encoding map. A common solution for this issue is to keep the total mass aside and use it when need to invert the encoding map. Therefore, only the mass distribution will be used to calculate the data misfit. For example, one can map u to $(\tilde{u}/\langle\tilde{u}\rangle, \langle\tilde{u}\rangle)$ with $\tilde{u} = \log(1 + \exp(u))$, and use the first element only for the misfit calculation; the second element is needed when inverting the map. In the following sections, encoding mappings that meet the above three conditions will be referred to as regular mappings.

5.1. Uniqueness of the gradient in the adjoint state method

According to theorem 4.3, for any $\mu \in \mathcal{U}$, the first variation of the transportation cost exists and is unique almost everywhere up to additive constants. Apparently, for all regular encoding maps, one expects that the gradient $\mathrm{d}\mathcal{J}/\mathrm{d}m$ in the adjoint state method does not depend on the particular choice of the Kantorovich potential φ. The following theorem presents a rigorous proof of this result.

Theorem 5.1. *Let m be a parameter model and u the data associated with m as in (8). For any fixed $\nu \in \mathcal{U}$, the value of*

$$\frac{\mathrm{d}}{\mathrm{d}m}\mathcal{T}(\mathcal{D}(u(m)), \nu) \tag{36}$$

does not depend on the particular law by which the Kantorovich potential φ is chosen, provided that $\mathcal{D}$ is differentiable.

Proof. Let m_0, m_1 be the first variations of $\mathcal{T}$ obtained with the particular choice of the Kantorovich potential, say φ_0 and φ_1, respectively. Recall from the adjoint state method (19) and (20), that m_k's are of the form

$$m_k = \int_0^T v_k \partial_t^2 u \,\mathrm{d}t, \quad k = 1, 2, \tag{37}$$

where u is the background wavefield, and v_k solves the adjoint wave equation with $\mathcal{D}'[u]^*(\varphi_k)$ as the right-hand side:

$$\begin{cases} \left(m(x)\dfrac{\partial_2}{\partial t^2} - \Delta\right) v_k(x,t) &= \mathcal{D}'[u]^*(\varphi_k), \\ v_k(x,T) &= 0, \\ \partial_t v_k(x,T) &= 0. \end{cases} \tag{38}$$

By theorem 4.3, we find that

$$\varphi_0 - \varphi_1 = c \tag{39}$$

for some constant c. We claim that

$$\int_\Omega (m_0 - m_1) h_m \,\mathrm{d}x = 0, \quad \forall\ h_m \in L_2. \tag{40}$$

To prove this, we consider an auxiliary wavefield h_u that solves the wave equation with $-h_m \partial_t^2 u$ as the right-hand side:

$$\begin{cases} \left(m(x)\dfrac{\partial_2}{\partial t^2} - \Delta \right) h_u(x,t) &= -h_m \partial_t^2 u, \\ h_u(x,0) &= 0, \\ \partial_t h_u(x,0) &= 0. \end{cases} \tag{41}$$

Then, it follows that

$$\begin{aligned} \int_\Omega (m_0 - m_1) h_m \,\mathrm{d}x &= \int_{\mathbb{R}^n} \left(\int_0^T (v_0 - v_1) \partial_t^2 u \,\mathrm{d}t \right) h_m \,\mathrm{d}x \\ &= -\int_{\mathbb{R}^n} \int_0^T (v_0 - v_1) \left(m(x)\frac{\partial_2}{\partial t^2} - \Delta \right) h_u(x,t) \mathrm{d}t \,\mathrm{d}x \\ &= -\int_{\mathbb{R}^n} \int_0^T h_u \left(\mathcal{D}'[u]^* \varphi_0 - \mathcal{D}'[u]^* \varphi_1 \right) \mathrm{d}t \,\mathrm{d}x \\ &= -\int_{\mathbb{R}^n} \int_0^T c\mathcal{D}'[u](h_u) \mathrm{d}t \,\mathrm{d}x \\ &= 0. \end{aligned} \tag{42}$$

In the above derivation, the first equality is from (37); substituting for $h_m \partial_t^2 u$ using (41), we obtain the second equality; the third equality employs (38) and integration by parts twice; then, we use the definition of the adjoint operator and (39) to conclude the proof. □

5.2. Encoding with softplus function

We now turn to the formulation of an encoding map using the softplus function. The Logistic function is defined as

$$f(x) = \frac{L}{1 + \mathrm{e}^{-\beta(x - x_0)}},$$

where x_0 is the value of the sigmoid's midpoint, L is the curve's maximum value, and β is the steepness of the curve. The standard logistic function is the one with parameters ($\beta = 1, x_0 = 0, L = 1$), which yields

$$f(x) = \frac{\mathrm{e}^x}{\mathrm{e}^x + 1} = \frac{1}{1 + \mathrm{e}^{-x}}.$$

The logistic function is useful since it can take any real number, whereas the output always takes values between zero and one and hence is interpretable as a probability density function. In practice, due to the nature of the exponential function e^{-x}, it is often sufficient to compute the standard logistic function for x over a small range of real numbers, such as a range contained in $[-5, 5]$. The anti-derivative of the logistic function $f(x) = \log(1 + \mathrm{e}^x)$ is widely used in logistic

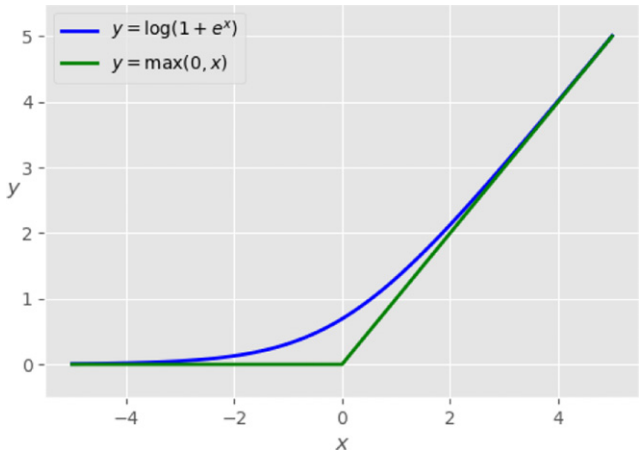


Figure 1. Softplus function and projection to the positive part.

regression, which is used in various areas, including machine learning and social sciences. The output also takes a positive value. Its derivative shows that the variance for negative input value is small. The graph of the function (figure 1) shows that the behavior of $f(x)$ is flat when $x < 0$ and is similar to $y = x$ when x is large.

We use the following operation

$$\tilde{u}(t) = \frac{1}{|\beta|} \log(1 + \mathrm{e}^{\beta u(t)}) \tag{43}$$

composited with the normalization

$$\tilde{u} \mapsto \frac{\tilde{u}}{\langle \tilde{u} \rangle}$$

to encode the seismic data into PDFs, where $\langle \cdot \rangle$ denotes the averaging operation. It is easy to check that

$$\lim_{\beta \to +\infty} \tilde{u} = u^+ \triangleq \max(u, 0) \quad \text{and} \quad \lim_{\beta \to -\infty} \tilde{u} = u^- \triangleq \max(-u, 0), \tag{44}$$

and the convergence is uniform. The above asymptotic behavior is an important advantage of this encoding procedure. One can expect the Wasserstein distance of the functions processed using this differentiable encoding method to show similar behavior as the one using u^+ while the smoothness preserved. According to the stability of the optimal transport plans [28, corollary 5.23], we can identify the convex functional on the seismic data u by checking its convexity on u^+ and u^-. In practice, large β can be chosen for better convexity in the objective function, but care should be taken to avoid the gradient-vanishing problem and overflow errors.

5.3. Convexity of the encoded data

We conclude this section by examining convexity under different measurement methods. The primary motivation for using the OT metric in the seismic inversion is to exploit its convexity to the translation and dilation, which are the primary data mismatch types. In [14], it is

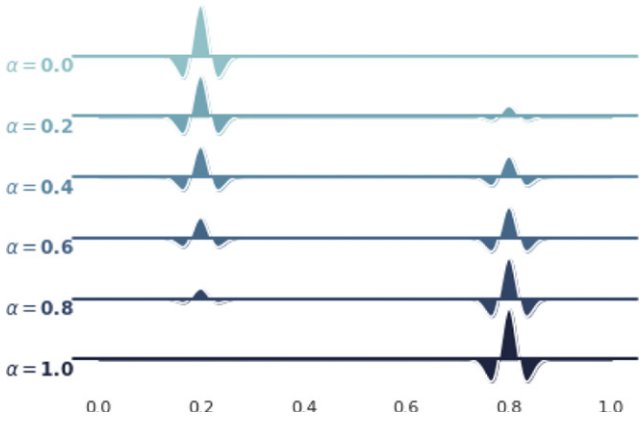


Figure 2. $\arg\min_p \left((1-\alpha)\|p-p_0\|_2^2 + \alpha\|p-p_1\|_2^2\right)$.

proved that the quadratic Wasserstein distance is convex with respect to translation and dilation, even in the case of a mixture of the two. In general, this convexity cannot be preserved after encoding. Roughly speaking, the encoding map can be interpreted as a procedure to generate non-negative functions from seismic data via adding/removing mass pointwise. After encoding, the endpoint, $t = 0, T$, can be a source or sink of mass. Hence the transportation cost is no longer convex to the translation and dilation.

Using the properties in (44) and [14, theorems 2.1–2.3], one can easily show that the encoded data using softplus function bears the asymptotic convexity when the pre-encoding data has compact support. Figures 2–4 present the interpolations of a Ricker wavelet $p_0(t)$ and its translation $p_1(t) = p_0(t-0.6)$ in L_2, W_2 with linear encoding method, and W_2 with softplus encoding, respectively. Unsurprisingly, the L_2 one calculates the interpolation in a pointwise manner; the encoding method using added constants shows a phenomenon of local transportation; by contrast, the one using softplus function accurately captures the translation information. The exponential encoding method [15], $\tilde{p}(t) = \mathrm{e}^{\alpha p(t)}$, can also retains the translation convexity with a large α, but a large α also leads to an overflow issue and it does not have an asymptotic behavior as $\alpha \to \pm\infty$.

6. Numerical examples

In this section, the properties of our proposed algorithm are illustrated through two numerical experiments. We first use simple structural models to investigate the relationship between the convexity of the misfit function and the encoding parameter β. The numerical experiment indicates that one can tune β to alleviate the local minima problem. Then, an inversion is performed on the 2D benchmark Marmousi model [36] to demonstrate the effectiveness of our method. To take advantage of the 1D explicit solution and avoid confusion on transportation over different units (time and space), instead of multiple-dimensional OT, we employ a trace-by-trace strategy to compute the objective function and the adjoint source. That is, we use the objective function

$$\mathcal{J}(u(x,t), \mathrm{d}(x,t)) = \int \mathcal{T}(u(x,t), \mathrm{d}(x,t))\mathrm{d}x.$$

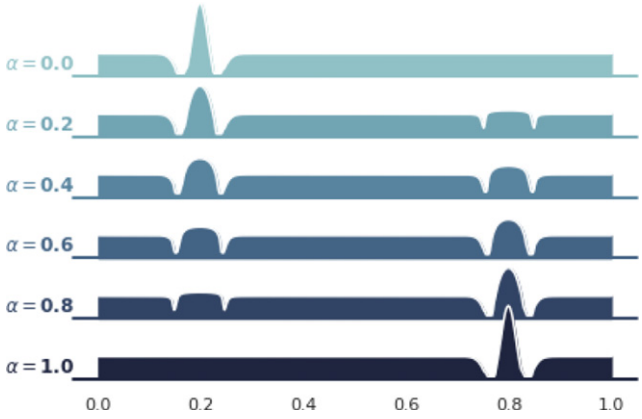


Figure 3. $\arg\min_p \left((1-\alpha)\mathcal{T}\left(p, \frac{p_0+c}{\langle p_0+c\rangle}\right) + \alpha\mathcal{T}\left(p, \frac{p_1+c}{\langle p_1+c\rangle}\right)\right)$.

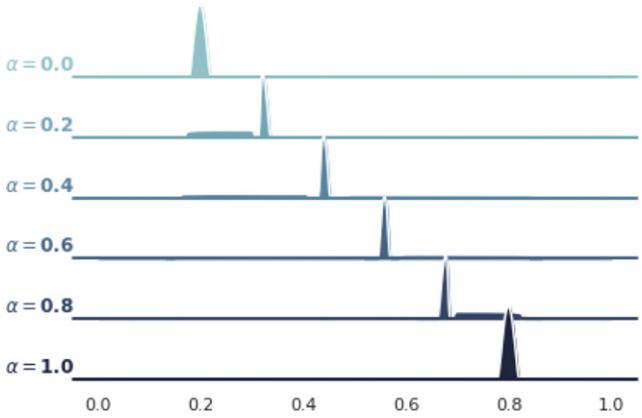


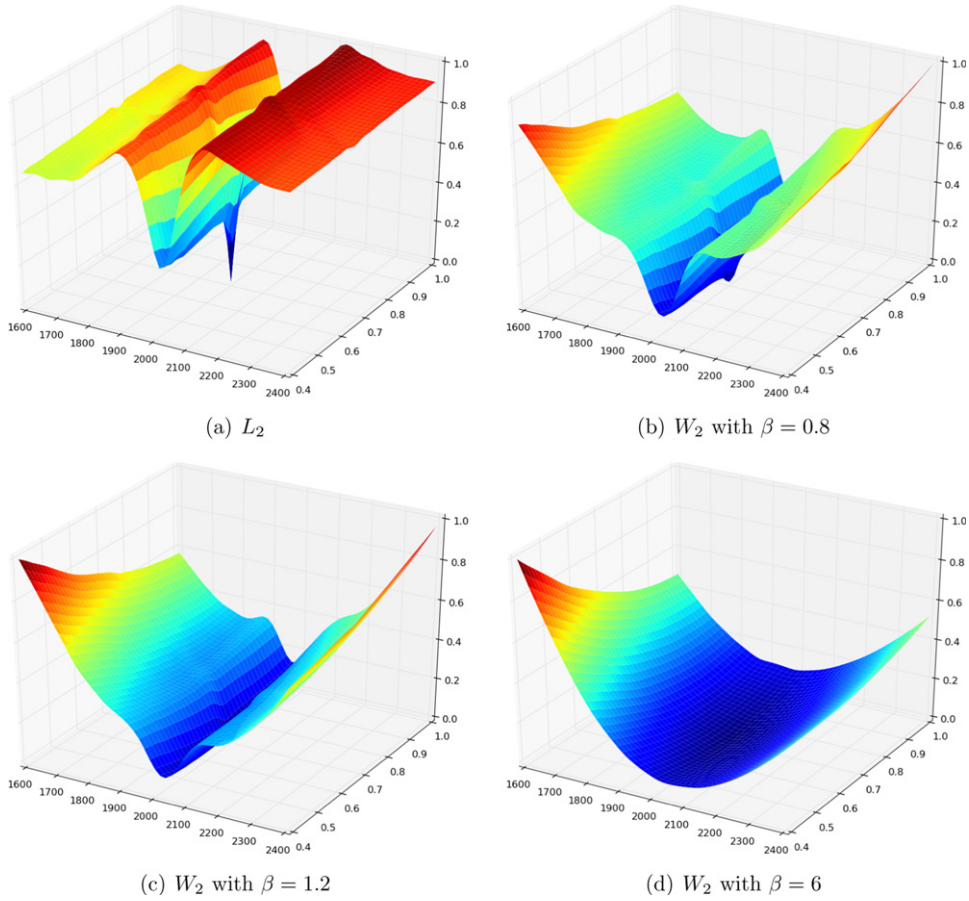
Figure 4. $\arg\min_p \left((1-\alpha)\mathcal{T}(p, \mathcal{D}(p_0)) + \alpha\mathcal{T}(p, \mathcal{D}(p_1))\right)$, $\mathcal{D}(p) = \frac{\log(1+e^p)}{\langle\log(1+e^p)\rangle}$.

6.1. The landscape of objective functions

We start our study of numerical experiments with a numerical investigation of the landscape of the misfit function. The experiment is performed on a family of 2D models with two feature variables. The receivers are uniformly distributed at an interval of 40 m over the top surface with 16.85 km length, and a point source is located in the center of the receivers. We use the following formula to build the velocity models:

$$v(x,z) = \begin{cases} 1500, & \text{when } z < 50, \\ v_0 + \alpha z, & \text{when } z \geqslant 50. \end{cases} \tag{45}$$

A band-pass filter at 3–18 Hz is applied to the source function and the data to imitate the actual exploration seismic data. The reference data is obtained with velocity model constructed with $v_0 = 2000$ m s^{-1} and $\alpha = 0.7$ s^{-1}. Figure 5 shows the misfit functions as functions of v_0 and α. For better comparison, we normalize the misfit using its maximum value.



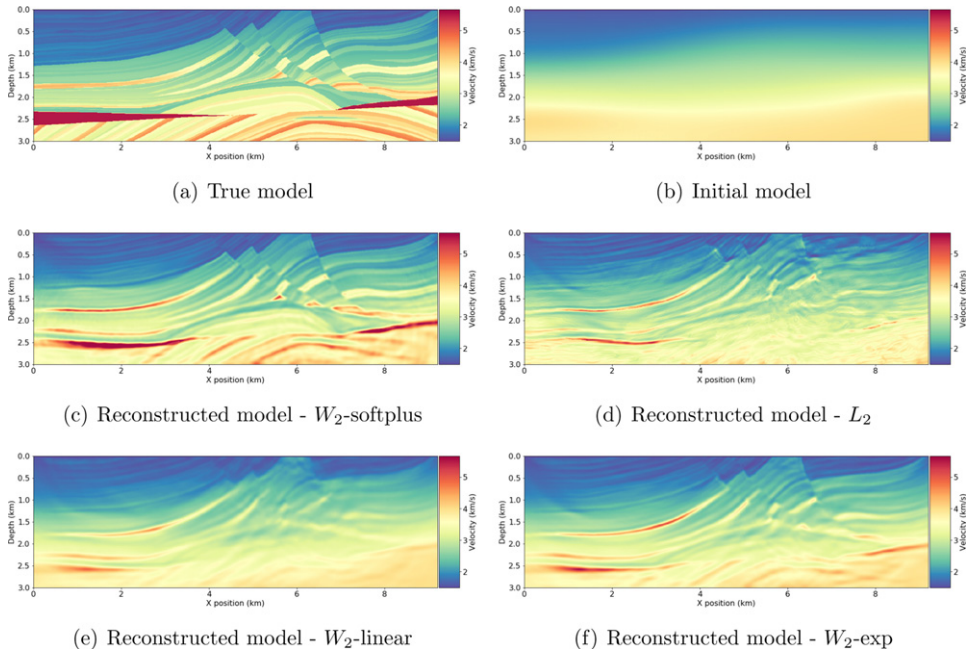
(a) L_2 (b) W_2 with $\beta = 0.8$

(c) W_2 with $\beta = 1.2$ (d) W_2 with $\beta = 6$

Figure 5. Comparison of landscapes.

The landscape using L_2 metric is shown in figure 5(a). Due to the high nonlinearity of the inverse problem and limited acquisition geometry, there are many local minima even for this simple-structured model. Once the background velocity is too far from the true model, the gradient of the misfit function will not update the model or even update it incorrectly. To arrive at the global minimum using a gradient-based descent method, one needs to start from an initial model within the same basin as the global minimum.

We further investigate the applicability of convexifying the W_2 misfit using encoding parameter β. This experimental setting provides a perfect scenario for the quadratic Wasserstein metric, since the number of the seismic events stays the same. Actually, it is easy to prove the asymptotic convexity of the objective function rigorously. Therefore, our goal is to eliminate the local minima by tuning β. Figures 5(b)–(d) displays the landscapes with gradually increasing β. It demonstrates that the larger β, the less local minima. It is also interesting to note that W_2 misfit functions are smoother than the L_2 one, which is associated with the fact that the regularity of the OT map is one degree higher than that of the seismic data.

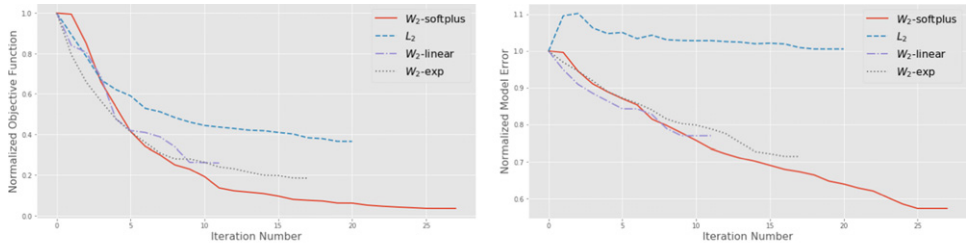


(a) True model (b) Initial model

(c) Reconstructed model - W_2-softplus (d) Reconstructed model - L_2

(e) Reconstructed model - W_2-linear (f) Reconstructed model - W_2-exp

Figure 6. Velocity models in the numerical experiment with Marmousi model.



(a) Objective function (b) Reconstruction error

Figure 7. Data and model error.

6.2. Inversion on the Marmousi model

In the following experiments, we use the Marmousi benchmark model [36]. The true velocity model is shown in figure 6(a). A 921 $\times$ 301 grid is used to represent an approximately 9.2 km $\times$ 3 km area. In both L_2 and W_2 cases, a heavily smoothed model from the true one, as shown in figure 6(b), is used as the initial model for the iterative gradient-based descent method (figure 7).

In this experiment, a perfectly matched layer absorbing boundary condition is applied to the domain boundaries except for the top free surface. The synthetic data is generated with an array of equally spaced 201 sources at depth 8 m and 461 receivers at depth 12 m distributed over the model's top surface. The source signature is the Ricker wavelet with a center frequency of 10 Hz, and the recording time is 4.5 s. A 3–18 Hz band-pass filter is applied to the source and

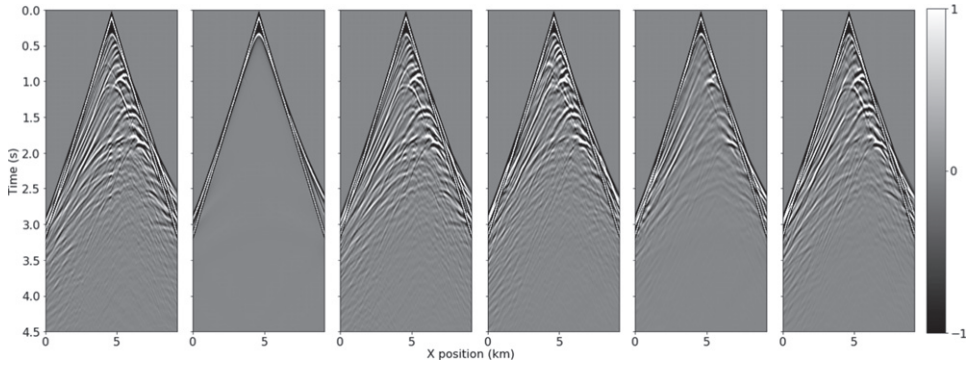


Figure 8. The single shot data generated with the corresponding models at $x = 4.6$ km. From left to right: true, initial, W_2-softplus, L_2, W_2-linear and W_2-exp.

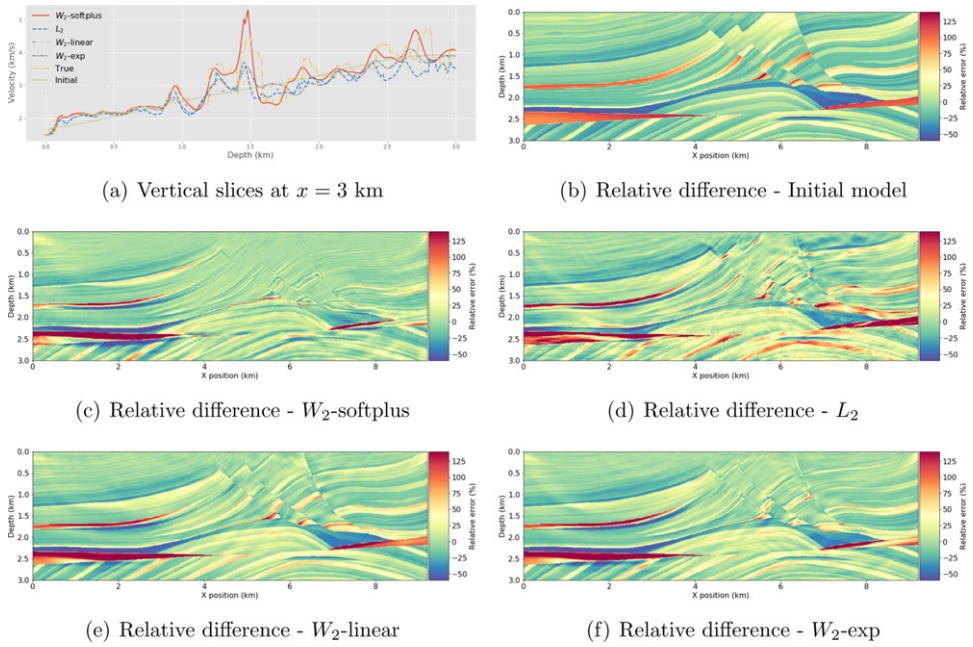
the data to imitate the actual seismic data in geophysical exploration. For the modeling and inversion, we use Devito [37] to solve the acoustic wave equation and the associated adjoint state equation. The numerical solution is obtained with a finite-difference scheme, which is forth-order accurate in space and second-order in time. We employ the limited-memory BFGS method with box constraints [38] implemented in SciPy [39] for the optimization. All the numerical experiments stop when the decrease of the objective function meets the stopping criteria,

$$\frac{\mathcal{J}_k - \mathcal{J}_{k+1}}{\max(\mathcal{J}_k, \mathcal{J}_{k+1}, 1)} < 10^{-5}.$$

Besides the proposed method, we also implement W_2 inversion with the linear [14] and exponential [15] encoding methods for comparison. The inversions using L_2 and W_2 with linear, exponential and softplus encoding stop after 20, 11, 17 and 27 iterations, respectively. The reconstruction results are displayed in figure 6, and the associated data are shown in figure 8. Due to the significant difference between the initial and true models, the least-squares formulation suffers from a cycle-skipping issue. The inversion using L_2 metric terminates with an incorrect velocity model. The W_2-linear one also stops early because of the local transportation issue; the problem of the exponential one is mainly in numerical aspects. The mainstream FWI solvers are all implemented using the single-precision floating-point accuracy for feasible memory usage, which prevents us from using large scaling parameters to improve convexity. At the same time, the softplus one can avoid the overflow issue by using spline approximations.

We present a vertical slice at $x = 3$ km in figure 9(a). By contrast, the L_2 metric produces low-velocity artifacts, which is strong evidence of cycle-skipping. Hot spots of slowness errors are shown in figure 9. The inversion with the proposed method correctly reconstructs the area swept by the diving waves. Some finer structures in the deeper region, mainly reflectors, can be improved using further iterations with a L_2 metric. The analysis in theorem 4.1 suggests that L_2 should be better for the inversion of details when it does not suffer from the cycle-skipping issue anymore. One can switch to L_2 metric once the cycle-skipping problem is overcome.

A fine-tuning procedure is usually not required for selecting β. A heuristic rule is to choose β to ensure that the total mass of the positive part is 50% greater than the total mass of the negative part after encoding if one wants to enhance the positive part, and vice versa. In this



(a) Vertical slices at $x = 3$ km

(b) Relative difference - Initial model

(c) Relative difference - W_2-softplus

(d) Relative difference - L_2

(e) Relative difference - W_2-linear

(f) Relative difference - W_2-exp

Figure 9. Vertical slices and relative slowness difference $(m - m_{\text{true}})/m_{\text{true}}$.

way, we can avoid local transportation and retain the translation convexity. A simple calculation shows that for the Ricker source wavelets with a maximum value of 1, the appropriate choice of β is 2.0, which is used in all inversion tasks.

7. Conclusion

We investigated the properties of the objective function for FWI using the quadratic Wasserstein metric and proper encoding methods. We rigorously prove that the quantity $\mathrm{d}\mathcal{J}/\mathrm{d}m$, obtained using the adjoint state method, does not depend on the particular choice of the Kantorovich potential if one chooses the encoding method properly. In particular, transportation metric with softplus encoding has asymptotic convexity concerning time-shift and dilation. It helps one extract time-shift information more accurately, thus provides the velocity model with appropriate large-scale changes, and mitigates the cycle-skipping problem.

For efficiency, we use a trace-by-trace approach in this work. FWI is a very computationally intensive task, and we do not want to add too much computational cost to it. One-dimensional OT has a closed-form solution and can be solved with linear complexity. Our approach only adds negligible cost to FWI (less than 0.1%).

Another point that should be stressed is that, based on the result in theorem 4.1, the transportation type objective function enhances low-frequency information as $\dot{H}^{-1}$ does. Thus, W_2 is more appropriate when the initial model is far from the true model. Once the cycle-skipping issue is fixed, it is better to switch to a L_2 metric for fast high-resolution reconstruction.

In two numerical examples, we show the feasibility of the proposed method. The first one uses two-parameter models to illustrate how the softplus encoding parameter recasts the landscape of the objective function. In the second example, we demonstrate the accuracy and efficiency of our method when applied to synthetic data generated by the Marmousi model.

We realize that a subtle choice of encoding parameter is not required. A heuristic rule is provided to choose β that will fit most cases. It is worth mentioning that some examples with field data show that either the positive or the negative part only is not enough due to ambiguity on the source wavelet. In this case, a sign-switch β can be applied to emphasizing positive and negative in an alternating fashion to improve the robustness of the proposed method.

As we mainly focus on applying the transportation metric on seismic data, optimization techniques only involving first-order derivatives are adopted. The Wasserstein metric with softplus encoding can be extended to be suitable for Newton's method or other second-order algorithms. Moreover, other than treating $\mathcal{T}(\cdot, \nu)$ as a function defined on L_2 and considering only the differential formulation in Euclidean sense, another natural strategy is to use Otto's calculus [28, chapter 15] and consider optimization using gradient flows in the Wasserstein space. We will investigate these approaches in future works.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 11971258). We are grateful to Björn Engquist and Yunan Yang for insightful discussions and comments. In particular, the author thanks the anonymous referees for their valuable comments, which helped improve the manuscript.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Appendix A. Algorithm

See (algorithm 1).

Algorithm 1. Calculation of $\mathcal{T}(p_0, p_1)$ and $\frac{\partial}{\partial p_1}\mathcal{T}(p_0, p_1)$.

procedure Pseudo-inverse(f_0, f_1, t) ▷ Calculation of $f_0^{-1}(f_1(t))$
  $m \leftarrow 1$
  for $k \leftarrow 1, N$ **do**
    $\tilde{f} = f_1(t_k)$
    while $f_0(t_m) < \tilde{f}$ **do** $m \leftarrow m + 1$
    if $m = 1$ **then** $\varphi(t_k) \leftarrow t_m$
    else if $m = N$ and $f(t_m) < \tilde{f}$ **then** $\varphi(t_k) \leftarrow t_m$
    else $\alpha \leftarrow \frac{\tilde{f} - f(t_{m-1})}{f(t_m) - f(t_{m-1})}$
      $\varphi(t_k) \leftarrow (1 - \alpha)t_{m-1} + \alpha t_m$ ▷ The function φ is $f_0^{-1}(f_1(t))$
procedure $\mathcal{T}(p_0, p_1, t)$ ▷ Calculation of $\mathcal{T}(p_0, p_1)$
  for $k \leftarrow 1, N$ **do**
    $f_0(t_k) \leftarrow f_0(t_{k-1}) + p_0(t_k), f_1(t_k) \leftarrow f_1(t_{k-1}) + p_1(t_k)$
  $\varphi \leftarrow$ Pseudo-inverse(f_0, f_1, t), $w \leftarrow 0$ ▷ The function φ is $f_0^{-1}(f_1(t))$
  for $k \leftarrow 1, N$ **do**
    $w \leftarrow w + p_1(t_k)\,(\varphi(t_k) - t_k)^2$ ▷ The value of $\mathcal{T}(p_0, p_1)$ is w
procedure gradient(p_0, p_1, t) ▷ Calculation of $\frac{\partial}{\partial p_1}\mathcal{T}(p_0, p_1)$
  for $k \leftarrow 1, N$ **do**
    $f_0(t_k) \leftarrow f_0(t_{k-1}) + p_0(t_k), f_1(t_k) \leftarrow f_1(t_{k-1}) + p_1(t_k)$
  $\varphi_0 \leftarrow$ Pseudo-inverse(f_0, f_1, t) ▷ φ_0 equals $f_0^{-1}(f_1(t))$
  $\varphi_1 \leftarrow$ Pseudo-inverse(f_1, f_0, t) ▷ φ_1 equals $f_1^{-1}(f_0(t))$
  $\xi \leftarrow$ Integration(φ_0, φ_1, t) ▷ $\xi(t)$ equals $\int_{f_0^{-1}(f_1(t))}^{1} \left(s - f_1^{-1}(f_0(s))\right) \mathrm{d}s$
  for $k \leftarrow 1, N$ **do**
    $\zeta(t_k) \leftarrow (\varphi_0(t_k) - t_k)^2 + 2\xi(t_k)$ ▷ ζ equals $\frac{\partial}{\partial p_1}\mathcal{T}(p_0, p_1)$

ORCID iDs

Lingyun Qiu https://orcid.org/0000-0002-2204-7235

References

[1] Albert T 1984 Inversion of seismic reflection data in the acoustic approximation *Geophysics* **49** 1259–66
[2] Plessix R-E 2006 A review of the adjoint-state method for computing the gradient of a functional with geophysical applications *Geophys. J. Int.* **167** 495–503
[3] Luo Y and Schuster G T 1991 Wave-equation traveltime inversion *Geophysics* **56** 645–53
[4] Van Leeuwen T and Mulder W A 2010 A correlation-based misfit criterion for wave-equation traveltime tomography *Geophys. J. Int.* **182** 1383–94
[5] Luo S and Paul S 2011 A deconvolution-based objective function for wave-equation inversion *SEG Technical Program Expanded Abstracts 2011* (Society of Exploration & Geophysicists) pp 2788–92
[6] Ma Y and Hale D 2013 Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion *Geophysics* **78** R223–33

[7] Warner M and Guasch L 2014 Adaptive waveform inversion-FWI without cycle skipping-theory *76th EAGE Conf. Exhibition 2014* (European Association of Geoscientists & Engineers) pp 1–5
[8] Symes W W 2015 Algorithmic aspects of extended waveform inversion *77th EAGE Conf. Exhibition-Workshops* vol 2015 (European Association of Geoscientists & Engineers) pp 1–5
[9] Symes W W 2017 Extended waveform inversion *79th EAGE Conf. Exhibition-Workshops* (European Association of Geoscientists & Engineers)
[10] Van Leeuwen T and Herrmann F J 2013 Mitigating local minima in full-waveform inversion by expanding the search space *Geophys. J. Int.* **195** 661–7
[11] van Leeuwen T and Herrmann F J 2015 A penalty method for PDE-constrained optimization in inverse problems *Inverse Problems* **32** 015007
[12] Wang C, Yingst D, Farmer P and Leveille J 2016 Full-waveform inversion with the reconstructed wavefield method *SEG Technical Program Expanded Abstracts 2016* (Society of Exploration Geophysicists) pp 1237–41
[13] Engquist B and Froese B D 2014 Application of the Wasserstein metric to seismic signals *Commun. Math. Sci.* **12** 979–88
[14] Engquist B, Froese B D and Yang Y 2016 Optimal transport for seismic full waveform inversion *Commun. Math. Sci.* **14** 2309–30
[15] Qiu L, Ramos-Martínez J, Valenciano A, Yang Y and Engquist B 2017 Full-waveform inversion with an exponentially encoded optimal-transport norm *SEG Technical Program Expanded Abstracts 2017* (Society of Exploration Geophysicists) pp 1286–90
[16] Qiu L, Ramos-Martínez J, Valenciano A, Jan K and Chemingui N 2017 Mitigating the cycle-skipping of full-waveform inversion: an optimal transport approach with exponential encoding *SEG 2017 Workshop: Full-waveform Inversion and Beyond* (Beijing, China 20–22 November 2017) (Society of Exploration Geophysicists) pp 1–4
[17] Yang Y and Engquist B 2018 Analysis of optimal transport and related misfit functions in full-waveform inversion *Geophysics* **83** 7–12
[18] 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** R43–62
[19] Engquist B and Yang Y 2019 Seismic imaging and optimal transport *Commun. Inf. Syst.* **19** 95–145
[20] Métivier L, Brossier R, Mérigot Q and Oudet E 2019 A graph space optimal transport distance as a generalization of L^p distances: application to a seismic imaging inverse problem *Inverse Problems* **35** 085001
[21] Métivier L, Brossier R, Mérigot Q, Oudet E and Virieux J 2016 An optimal transport approach for seismic tomography: application to 3D full waveform inversion *Inverse Problems* **32** 115008
[22] Poncet R, Messud J, Bader M, Lambaré G, Viguier G and Hidalgo C 2018 FWI with optimal transport: a 3D implementation and an application on a field dataset *80th EAGE Conf. Exhibition 2018* (European Association of Geoscientists & Engineers) pp 1–5
[23] Villani C 2003 *Topics in Optimal Transportation* (*Graduate Studies in Mathematics* vol 58) (Providence, RI: American Mathematical Society)
[24] Bogachev V I 2007 *Measure Theory* (Berlin: Springer)
[25] Engquist B, Ren K and Yang Y 2020 The quadratic Wasserstein metric for inverse data matching *Inverse Problems* **36** 055001
[26] Li D, Lamoureux M and Liao W 2020 Full waveform inversion with unbalanced optimal transport distance (arXiv:2004.05237)
[27] Sun B and Alkhalifah T 2019 The application of an optimal transport to a preconditioned data matching function for robust waveform inversion *Geophysics* **84** R923–45
[28] Villani C 2008 *Optimal Transport: Old and New* vol 338 (Berlin: Springer)
[29] Santambrogio F 2015 *Optimal Transport for Applied Mathematicians* (Basel: Birhäuser)
[30] Ambrosio L, Gigli N and Savaré G 2008 *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Berlin: Springer)
[31] Bao G, Li P, Lin J and Triki F 2015 Inverse scattering problems with multi-frequencies *Inverse Problems* **31** 093001
[32] de Hoop M V, Qiu L and Scherzer O 2015 An analysis of a multi-level projected steepest descent iteration for nonlinear inverse problems in Banach spaces subject to stability constraints *Numer. Math.* **129** 127–48
[33] Carey B, Saleck F M, Zaleski S and Chavent G 1995 Multiscale seismic waveform inversion *Geophysics* **60** 1457–73

[34] Jean V and Operto S 2009 An overview of full-waveform inversion in exploration geophysics *Geophysics* **74** WCC1–26
[35] Fichtner A, Trampert J, Cupillard P, Saygin E, Taymaz T, Capdeville Y and Villaseñor A 2013 Multiscale full waveform inversion *Geophys. J. Int.* **194** 534–56
[36] Versteeg R 1994 The marmousi experience: velocity model determination on a synthetic complex data set *Lead. Edge* **13** 927–36
[37] Lange M, Kukreja N, Louboutin M, Luporini F, Vieira F, Pandolfo V, Velesko P, Kazakas P and Gorman G 2016 Devito: towards a generic finite difference DSL using symbolic python *2016 6th Workshop on Python for High-Performance and Scientific Computing (PyHPC)* pp 67–75
[38] Byrd R H, Lu P, Nocedal J and Zhu C 1995 A limited memory algorithm for bound constrained optimization *SIAM J. Sci. Comput.* **16** 1190–208
[39] Virtanen P and Gommer R (SciPy 1.0 Contributors) 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72