.

OPTIMAL TRANSPORT FOR SEISMIC FULL WAVEFORM INVERSION

BJÖRN ENGQUIST, BRITTANY D. FROESE, AND YUNAN YANG

Abstract. Full waveform inversion is a successful procedure for determining properties of the earth from surface measurements in seismology. This inverse problem is solved by a PDE constrained optimization where unknown coefficients in a computed wavefield are adjusted to minimize the mismatch with the measured data. We propose using the Wasserstein metric, which is related to optimal transport, for measuring this mismatch. Several advantageous properties are proved with regards to convexity of the objective function and robustness with respect to noise. The Wasserstein metric is computed by solving a Monge-Ampère equation. We describe an algorithm for computing its Frechet gradient for use in the optimization. Numerical examples are given.

1. Introduction

A central step in seismic exploration is the estimation of basic geophysical properties. This can, for example, be wave velocity, which is what we will consider here. This step is typically the basis of an imaging process to determine geophysical structures.

The computational technique full waveform inversion (FWI) was introduced to seismology in [11, 20]. This inverse method follows the common strategy of PDE constrained optimization. The unknown wave velocity v is determined by minimizing the mismatch $d(f, g)$ between the simulated wave field $f(v)$ and the measured data g.

$$v^* = \operatorname*{argmin}_v F(v), \quad F(v) = d(f(v), g) + \lambda R(v).$$

Since many forms of the inverse problem are ill-posed [15, 17, 21], some regularization $R(v)$ is widely used to ensure the uniqueness of model recovery. We will here focus on properties of misfit measure d and not on the overall question of uniqueness and stability of the inverse problem. For example, see [9, 18, 19] for theories of uniqueness and stability.

The L_2 norm is often used to measure the misfit, which typically generates many local minima. This problem is exacerbated by the fact that measured signals usually suffer from noise in the measurements [12].

In [6], we proposed using the Wasserstein metric for the misfit function, i.e. $d(f, g) = W_2^2(f, g)$. The Wasserstein metric measures the distance between two distributions as the optimal cost of rearranging one distribution into the other [22]. The mathematical definition of the distance between the distributions $f : X \to \mathbb{R}^+$,

Date: February 5, 2016.

The research were supported by the Texas Consortium for Computational Seismology. The first and the third author were also partially supported by NSF grant DMS-1522792.

$g : Y \to \mathbb{R}^+$ can be formulated as

$$W_2^2(f,g) = \inf_{T \in \mathcal{M}} \int_X |x - T(x)|^2 f(x)\, dx \tag{1}$$

where $\mathcal{M}$ is the set of all maps that rearrange the distribution f into g. The optimal transport formulation requires nonnegative distributions. To define a distance between more general f and g we can, for example, replace $W_2^2(f,g)$ by $W_2^2(f^+,g^+) + W_2^2(f^-,g^-)$ where $f^+ = \max\{f,0\}$, $f^- = \max\{-f,0\}$.

It is our goal to prove several desirable properties relating to convexity and insensitivity to noise, which were briefly discussed in [6]. Another important contribution in this paper is a derivation of the gradient of $d(f(v),g)$ with respect to v, which is essential for gradient based minimization algorithms. It is outside the scope of this work to study serious applications, but we give some numerical examples to show the quantitative behavior and to compare with the simple search algorithm used in [6]. In this earlier paper, simple geometrical optics was used in the forward problem. Here we consider the full wave equation.

We briefly recall one example from [6] that illustrates the advantage of the Wasserstein metric. Consider the misfit between the simple wavelet f in Figure 1(a) and another wavelet shifted by a distance s. Figures 1(b)-1(c) illustrate that the L_2 norm is constant when s is large and has many local minima. On the other hand, the Wasserstein metric is uniformly convex with respect to shifts, which are natural in travel time mismatches.

Earlier algorithms for the numerical computation of the Wasserstein metric required a large number of operations [1, 3, 4]. The optimal transportation problem can be rigorously related to the following Monge-Ampère equation [5, 10], which enables the construction of more efficient methods for computing the Wasserstein metric.

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

The Wasserstein metric is then given by

$$W_2^2(f,g) = \int_X |x - \nabla u(x)|^2 f(x)\, dx. \tag{3}$$

There are now fast and robust numerical algorithms for the solution of (2), and thus for the computation of W_2^2, and these form the basis for our numerical techniques [2].

In section 2, we will study the convexity of the quadratic Wasserstein metric with respect to shift, dilation, and partial amplitude change. Error between simulated data and observed data in the form of shifts and dilations can occur naturally from an incorrect velocity model parameter, while inaccurate measurements on the surface can result in larger or smaller local amplitudes. We will give a rigorous proof of these convexity statements using the fundamental theorem of optimal transport and convexity of the Monge-Kantorovich minimization problem [22].

In section 3, we will discuss how the Wasserstein metric is affected by random noise with a uniform distribution. For the optimal transport problem on the real line, both a theorem and numerical illustrations will be given to show that the

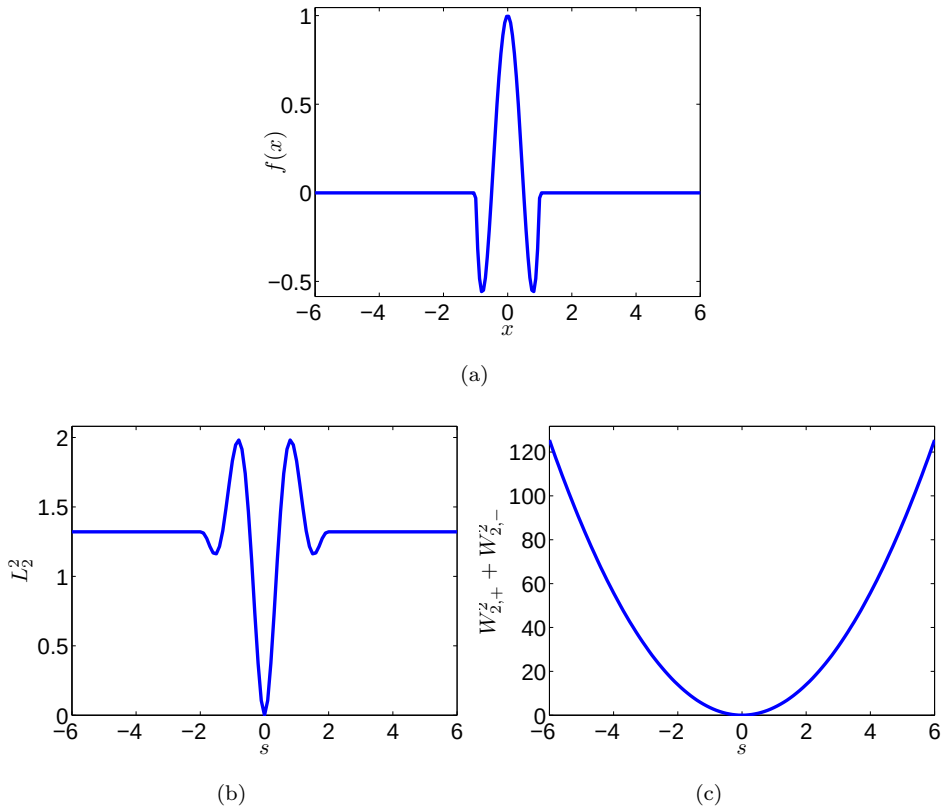


FIGURE 1. (a) A wavelet profile $f(x)$. The distance between $f(x)$ and $g(x) = f(x-s)$ measured by (b) $L_2^2(f,g)$, and (c) $W_2^2(f^+,g^+)+W_2^2(f^-,g^-)$[6].

effect of noise is negligible. For higher dimensions, we estimate the effect of noise by finding an upper bound.

We review an efficient numerical method for computing the Wasserstein metric via the numerical solution of the nonlinear elliptic Monge-Ampère partial differential equation in section 4. After obtaining the discrete solution, we can easily approximate the squared Wasserstein metric.

We are interested in recovering the parameters in the forward wave equation by minimizing the Wasserstein metric between simulated and observed data. In section 5, we discuss two different ways to obtain the gradient of the Wasserstein metric. We can linearize the continuous problem first, then discretize the resulting linear elliptic equation. Alternatively, we can first discretise the Wasserstein metric, then linearise the result. This second approach is particularly straightforward for the numerical method utilized in this paper.

Finally, numerical examples presented in section 6 show the quantitative and qualitative behavior of the minimization procedure. Parameters in low dimensional model problems are recovered by minimizing the Wasserstein metric without regularization.

2. Convexity of the quadratic Wasserstein metric

In most optimization problems, convexity of the objective function is a desirable property. The example of convexity given in Figure 1(c) was our motivation for considering the Wasserstein metric in the context of full waveform inversion. In this section, we will mathematically study this convexity with respect to shift, dilation, and local change in amplitude.

We analyze cases where f is derived from g by either a local change of amplitude or a linear change of variables in the form of a shift or dilation. The change in amplitude may originate from measurement errors and variations in strength of reflecting surfaces. The shift and dilation are typical effects of variations in the velocity v, as can be seen in a simple example.

The starting point is the formulation as the PDE constrained optimization given in the introduction,

$$\min_v d(f(v), g) = \min_{v(x)} W_2^2\left(f(x,t;v), g(x,t)\right).$$

A simple one-dimensional, constant velocity model is

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = v^2 \frac{\partial^2 u}{\partial x^2}, & x > 0, t > 0, \\ u = 0, \quad \frac{\partial u}{\partial t} = 0, & x > 0, t = 0, \\ u = u_0(t), & x = 0, t > 0. \end{cases}$$

One solution to the equation is $u(x,t;v) = u_0(t - x/v)$. For fixed x, variations in v induce shifts in the signal. When t is fixed, variation of v generates dilation in u_0 as a function of x.

2.1. Convexity with respect to shift.

We assume the optimal map between two density functions f and g is T. Given $\eta \in X$, we define a new distribution $f_s : X \to \mathbb{R}$, $f_s(x) = f(x - s\eta)$. The corresponding optimal map between f_s and g is T_s. The relation between T and T_s is as follows:

Theorem 1 (Convexity of shift)**.** *Suppose f and g are density functions of two Borel probability measures with finite second moment. Let T be the optimal map that rearranges f into g. If $f_s(x) = f(x - s\eta)$ for $\eta \in X$, then the optimal map from $f_s(x)$ to $g(y)$ is $T_s = T(x - s\eta)$. Moreover, $W_2^2(f_s, g)$ is convex with respect to s.*

The proof includes several fundamental concepts and theorems from optimal transportation.

Definition 1 (Cyclical monotonicity)**.** *We say that $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^n$ is cyclically monotone if for any $m \in \mathbb{N}^+$, $(x_i, y_i) \in \Gamma, 1 \le i \le m$ implies that*

$$\sum_{i=1}^{m} |x_i - y_i|^2 \le \sum_{i=1}^{m} |x_i - y_{i-1}|^2 \tag{4}$$

or equivalently

$$\sum_{i=1}^{m} \langle y_i, x_i - x_{i-1} \rangle \ge 0 \tag{5}$$

where $x_0 \equiv x_m$ and $y_0 \equiv y_m$.

Given a Polish space (X, d) (i.e. a complete and separate metric space), we let $\mathscr{P}(X)$ be the set of Borel probability measures on X. The transport plan is given by a probability measure π on the product space $X \times Y$. The set $\Pi(\mu, \nu)$ consists of all transport plans $\pi \in \mathscr{P}(X \times Y)$ from μ to ν, i.e. the set of Borel probability measures on $X \times Y$ such that

$$\pi(A \times Y) = \mu(A) \quad \forall A \in \mathscr{B}(X), \qquad \pi(X \times B) = \nu(B) \quad \forall B \in \mathscr{B}(Y).$$

The next two lemmas from [14, 22] show the equivalence of optimality and cyclical monotonicity under the condition that μ does not give mass to small sets.

Lemma 2 (Optimality criterion for quadratic cost)**.** *Let μ and ν be two probability measures on $\mathbb{R}^N$, where μ does not give mass to small sets. Let $\pi \in \Pi(\mu, \nu)$ with cyclically monotone support. Then π is optimal in the Kantorovich problem of mass transference between μ and ν with quadratic cost $c(x, y) = |x - y|^2$.*

Lemma 3 (Optimal plans have cyclically monotone support)**.** *Let μ, ν be two probability measures on $\mathbb{R}^n$ and let $\pi \in \Pi(\mu, \nu)$ be optimal in the Kantorovich problem of mass transference between μ and ν with quadratic cost $c(x, y) = |x-y|^2$. Then the support of π is cyclically monotone.*

Proof of Theorem 1. By our original assumption, the optimal map between two measures with density functions f and g is T. We will show that the new joint measure $\pi_s = (Id \times T_s)\#\mu_s$ is cyclically monotone.

With $y_i = T_s(x_i)$ and $T_s(x) = T(x - s\eta)$, we need the following inequality to hold:

$$\begin{aligned} &\sum_{i=1}^{m} \langle y_i, x_i - x_{i-1} \rangle \geq 0 \\ \iff &\sum_{i=1}^{m} x_i \cdot T(x_i - s\eta) \geq \sum_{i=1}^{N} x_i \cdot T(x_{i-1} - s\eta) \\ \iff &\sum_{i=1}^{N} (x_i - s\eta) \cdot T(x_i - s\eta) \geq \sum_{i=1}^{m} (x_i - s\eta) \cdot T(x_{i-1} - s\eta) \end{aligned}$$

The last inequality is just a statement of cyclical monotonicity of the joint measure $\pi = (Id \times T)\#\mu$ for f and g without the shift. This is automatically true since by assumption T is the optimal in that setting.

By the uniqueness of monotone measure-preserving optimal maps between two distributions [13], we assert that $T_s(x) = T(x - s\eta)$ is the optimal map corresponding to the shifted function $f(x - s\eta)$.

Consequently, $W_2^2(f_s, g)$ is given by

$$\begin{aligned} W_2^2(f_s, g) &= \int |x - T_s(x)|^2 f_s(x) dx \\ &= \int |x - T(x - s\eta)|^2 f(x - s\eta) dx \\ &= W_2^2(f, g) + s^2|\eta|^2 + 2s \int < \eta, x - T(x) > f(x) dx. \end{aligned} \tag{6}$$

The convexity with respect to s is evident from the last equation. □

2.2. **Convexity with respect to dilation.**

Theorem 4 (Optimal map for dilation)**.** *Assume $g(y)$ is a density function of finite second moment and $f(x) = \det(A)g(Ax)$, where A is a symmetric positive definite matrix. Then the optimal transport map to rearrange $f(x)$ into $g(y)$ is $T(x) = Ax$.*

Proof. Again, Lemmas 2 and 3 are the tools to verify this optimal map. We will show that the joint measure $\pi = (Id \times T)\#\mu_f$ is cyclically monotone. It is easy to show that T is measure-preserving:

$$g(T(x))\det(\nabla T) = g(Ax)\det(A) = f(x) \tag{7}$$

Since A is symmetric positive definite, it has a unique Cholesky decomposition, i.e. $A = L^T L$ for some upper triangular matrix L. For any $m \in \mathbb{N}$:

$$\begin{aligned}
& \sum_{i=1}^{m} |x_i - T(x_i)|^2 - \sum_{i=1}^{m} |x_i - T(x_{i-1})|^2 \\
= \; & -2\sum_{i=1}^{m} (x_i^T A x_i - x_{i-1}^T A x_i) \\
= \; & -2\sum_{i=1}^{m} (|Lx_i|^2 - x_{i-1}^T L^T L x_i) \\
= \; & -\sum_{i=1}^{m} (|Lx_i|^2 - 2x_{i-1}^T L^T L x_i + |Lx_{i-1}|^2) \\
= \; & -\sum_{i=1}^{m} |Lx_i - Lx_{i-1}|^2 \leq 0,
\end{aligned}$$

which justifies equation (4). □

Remark 1. The requirement that A be symmetric positive definite is necessary for $y = Ax$ to be the optimal map. For example, let A be the rotation matrix $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ with $\theta = \pi$ and let $g(x, y) = g(-x, -y)$. The optimal map is the identity function instead of $T = A^{-1}x$.

Convexity is a separate question as it depends on the parameterisation. One special case of dilation occurs when A is a diagonal matrix. The following theorem is a natural consequence of the definition and Theorem 4.

Theorem 5 (Convexity with respect to dilation)**.** *Assume $g(y)$ is a density function of finite second moment and $f_\lambda(x) = \frac{1}{\lambda^n} g(\frac{x}{\lambda})$. Then the optimal map between f and g is $T(x) = \frac{x}{\lambda}$. The Wasserstein metric $W_2^2(f_\lambda, g)$ is $(1-\lambda)^2 \int y^2 g(y)dy$, a convex function of λ. More generally, if the dilation matrix $A = diag(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n})$ and $f(x) = \det(A)g(Ax)$, the Wasserstein metric $W_2^2(f, g)$ is convex with respect to $\lambda_1, \ldots, \lambda_n$.*

Remark 2. If both dilation and shift are present, the Wasserstein metric will be convex with respect to each of the corresponding parameters.

2.3. **Convexity with respect to partial amplitude change.** Now we consider density functions f_α and g on a domain $\Omega = \Omega_1 \cup \Omega_2$ with $\Omega_1 \cap \Omega_2 = \emptyset$. Assume function f_α depends on g as follows:

$$f_\alpha(x) = \begin{cases} (1+\alpha)g(x), & x \in \Omega_1, \\ (1-\gamma_\alpha)g(x), & x \in \Omega_2, \end{cases}$$

where

$$\gamma_\alpha = \alpha \frac{\int_{\Omega_1} g}{\int_{\Omega_2} g}.$$

Here α is related to the size of the "perturbation". Obviously, $f_0(x) = g(x)$ for any $x \in \Omega$. The rescaling parameter γ ensures both distributions have the same mass

$$\int_\Omega f_\alpha(x)dx = \int_\Omega g(x)dx,$$

which is necessary to evaluate the Wasserstein metric.

Theorem 6 (Convexity with respect to partial amplitude change)**.** *With the density functions f_α and g defined as above, the Wasserstein metric $W_2^2(f_\alpha, g)$ is a convex function of the parameter α.*

Proof. Choose any α_1, α_2 such that f_{α_1} and f_{α_2} are nonnegative, $s \in [0,1]$, and let h be an arbitrary density function. From convexity of the Monge-Kantorovich minimization problem [22], we have

$$W_2^2\left(h, sf_{\alpha_1} + (1-s)f_{\alpha_2}\right) \le sW_2^2\left(h, f_{\alpha_1}\right) + (1-s)\, W_2^2(h, f_{\alpha_2}). \tag{8}$$

We can calculate

$$\begin{aligned} sf_{\alpha_1} + (1-s)f_{\alpha_2} &= \begin{cases} s(1+\alpha_1)g + (1-s)(1+\alpha_2)g, & x \in \Omega_1, \\ s(1-\gamma_{\alpha_1})g + (1-s)(1-\gamma_{\alpha_2})g, & x \in \Omega_2. \end{cases} \\ &= \begin{cases} (1 + s\alpha_1 + \alpha_2 - s\alpha_2)g, & x \in \Omega_1, \\ (1-\gamma_{s\alpha_1+(1-s)\alpha_2})g, & x \in \Omega_2, \end{cases} \\ &= f_{s\alpha_1+(1-s)\alpha_2}. \end{aligned}$$

Thus we can rewrite Equation (8) as

$$W_2^2(h, f_{s\alpha_1+(1-s)\alpha_2}) \le sW_2^2(h, f_{\alpha_1}) + (1-s)W_2^2(h, f_{\alpha_2}) \tag{9}$$

and the Wasserstein metric $W_2^2(h, f_\alpha)$ is convex with respect to the amplitude change parameter α. □

3. Insensitivity with respect to noise

In the practical application of full waveform inversion, it is natural to experience noise in the measured signal, and therefore robustness with respect to noise is a desirable property in a measure of mismatch. We will show that Wasserstein metric is substantially less sensitive to noise than the L^2 norm.

The Wasserstein metric depends on the square of the translate T. This implies that if f is an oscillatory perturbation of g then the Wasserstein metric $W_2^2(f, g)$ is small. One example given by Villani [22] shows that $W_2^2(\mu_k, \mathcal{L}) = O(\epsilon^2)$ for $d\mu_k = \left(1 + \sin\frac{2\pi x}{\epsilon}\right) dx$ and the Lebesgue measure on $[0,1]$. A numerical example without analysis was given in [6].

3.1. **One dimension.** In one dimension, it is possible to exactly solve the optimal transportation problem in terms of the cumulative distribution functions

$$F(x) = \int_0^x f(t)\,dt, \quad G(x) = \int_0^x g(t)\,dt.$$

See Figure 2.

Lemma 7 (Optimal transportation for a quadratic cost on $\mathbb{R}$ [22])**.** *Let μ, ν be two probability measures on $\mathbb{R}$, with respective cumulative distribution functions F and G. Let π be the probability measure on $\mathbb{R}^2$ with joint two-dimensional cumulative distribution function*

$$H(x, y) = \min(F(x), G(y)). \tag{10}$$

Then, $\pi \in \Pi(\mu, \nu)$ is optimal in the Kantorovich transportation problem between μ and ν for the quadratic cost function $c(x, y) = |x - y|^2$. Moreover, the value of the optimal transportation cost is

$$W_2^2(\mu, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt. \tag{11}$$

Remark 3. If μ does not give mass to points, then $T = G^{-1} \circ F$ transports μ onto ν, and $\int_{-\infty}^{x} d\mu = \int_{-\infty}^{T(x)} d\nu$.

Theorem 8 (Insensitivity to noise in 1-D)**.** *Let ν be a probability measure on $[0, 1]$ with non-negative density g and choose $0 < c < \min g$. Let μ_N be the probability measure with density*

$$\frac{g(x) + r^N(x)}{1 + \bar{r^N}},$$

which contains piecewise constant additive noise

$$r^N(x) \equiv r_i, \quad x \in \left(\frac{i-1}{N}, \frac{i}{N}\right], \quad 1 \le i \le N$$

with each r_i drawn from the uniform distribution $U[-c, c]$. Then $\mathbb{E}W_2^2(\mu_N, \nu)$ is $\mathcal{O}(\frac{1}{N})$.

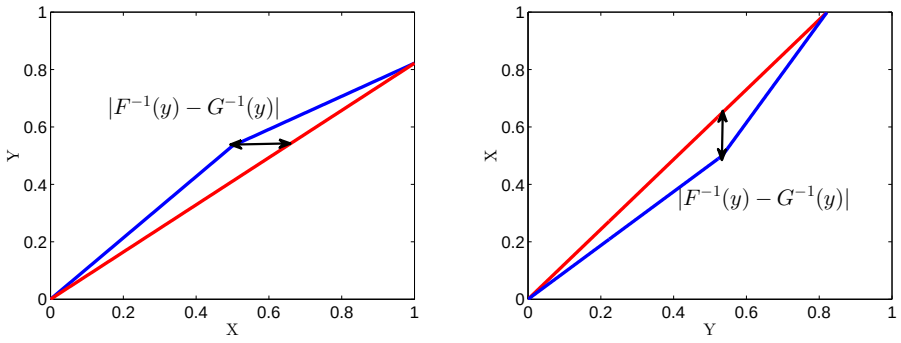
Without loss of generality, we take ν to be the Lebesgue measure on $[0, 1]$. Figure 3 shows the effect of noise. As $N \to \infty$, r^N approximates the noise function $r(x)$ on $[0, 1]$. For any $x_0 \in [0, 1]$, $r(x_0)$ is a random variable with uniform distribution on $[-c, c]$.

Proof of Theorem 8. Let ν be the Lebesgue measure on $[0, 1]$, with density function $g = 1$. For each i, r_i is a random variable of uniform distribution on $[-c, c]$, $0 < c < 1$. Thus, we have $\mathbb{E}r_i = 0$ and $\mathbb{E}\bar{r} = 0$.

Let $h = 1/N$ and $x_i = ih$ for $i = 0, \ldots, N$. Then the noisy density function is given by

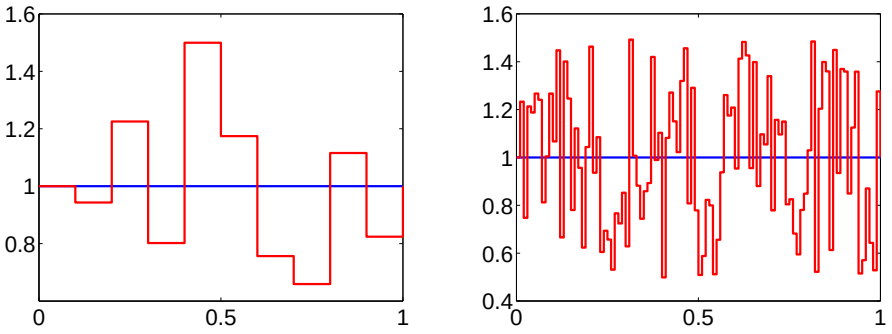
$$f_N(x) = 1 + r_i, \quad x \in (x_{i-1}, x_i].$$

We begin by calculating the Wasserstein metric between f_N and $g_N = 1 + \bar{r}^N$, which share the same mass.



(a) (b)

FIGURE 2. (a) Cumulative distribution functions $F(x)$ (red), $G(x)$ (blue) and (b) inverse functions $F^{-1}(y)$, $G^{-1}(y)$.



(a) (b)

FIGURE 3. Densities of ν (blue) and μ_N (red, $c = -0.5$) for (a) $N = 10$ and (b) $N = 100$.

According to Theorem 7, we need the cumulative distribution function and its inverse for both f_N and g_N. It is easy to derive

(12)
$$F_N(x) = \int_0^x f_N(x)dx = \begin{cases} (1+r_1)x, & x \in [0,h] \\ (1+r_1)h + (1+r_2)(x-h), & x \in (h,2h] \\ \quad\vdots & \\ \sum_{i=1}^{N-1}(1+r_i)h + (1+r_N)(x-(N-1)h), & x \in (1-h,1] \end{cases}$$

(13)
$$G_N(x) = \int_0^x g_N(x)dx = (1+\bar{r})x, \quad 0 \le x \le 1$$

$$F_N^{-1}(x) = \begin{cases} \frac{x}{1+r_1}, & x \in [0, (1+r_1)h] \\ \frac{x+(r_2-r_1)h}{1+r_2}, & x \in ((1+r_1)h, (2+r_1+r_2)h] \\ \vdots & \\ \frac{x+((N-1)r_N-\sum_{i=1}^{N-1} r_i)h}{1+r_N}, & x \in (\sum_{i=1}^{N-1}(1+r_i)h, 1+\bar{r}] \end{cases} \tag{14}$$

$$G_N^{-1}(x) = \frac{x}{1+\bar{r}}, \quad 0 \le x \le 1+\bar{r}. \tag{15}$$

Then we can estimate the squared Wasserstein metric by

$$\begin{aligned} W_2^2(f_N, g_N) &= \int_0^{1+\bar{r}} |F_N^{-1}(t) - G_N^{-1}(t)|^2 dt \\ &\le \max\left\{ \left(\frac{(1+r_1)h}{1+\bar{r}} - h \right)^2, 0 \right\} \cdot (1+r_1)h + \\ &\quad \max\left\{ \left(\frac{(1+r_1)h}{1+\bar{r}} - h \right)^2, \left(\frac{(2+r_1+r_2)h}{1+\bar{r}} - 2h \right)^2 \right\} \cdot (1+r_2)h + \cdots + \\ &\quad \max\left\{ (0, \left(\frac{(1+r_1)h + \cdots + (1+r_{N-1})h}{1+\bar{r}} - (N-1)h \right)^2 \right\} \cdot (1+r_N)h \\ &\le 2 \cdot \frac{h^3}{(1-c)^2} \left\{ (r_1 - \bar{r})^2 + (r_1 + r_2 - 2\bar{r})^2 + \cdots + (r_1 + \cdots + r_{N-1} - (N-1)\bar{r})^2 \right\} \\ &\le \frac{2h^3}{(1-c)^2} \sum_{i=1}^{N} \left(\sum_{l=1}^{i} r_l - ih \sum_{k=1}^{N} r_k \right)^2. \end{aligned}$$

Since the noise $\{r_i\}_{i=1}^N$ is i.i.d., we obtain the following upper bound for the expectation of the Wasserstein metric:

$$\mathbb{E}W_2^2(f_N, g_N) \le C \cdot h^3 \cdot \sum_{i=1}^{N} i \cdot \mathbb{E}r_1^2 \le \frac{C_2}{N}.$$

We can prove the lower bound $\mathbb{E}W_2^2(f_N, g_N) \ge \frac{C_1}{N}$ in a similar way. Thus

$$\frac{C_1}{N} \le \mathbb{E}W_2^2(f_N, g_N) \le \frac{C_2}{N} \tag{16}$$

where C_1 and C_2 only depend on c.

The density functions f_N and g_N have total mass $1+\bar{r}^N$, and must be rescaled in order to obtain the probability measures μ_N and ν. That is, the density of μ_N is $f_N/(1+\bar{r}^N)$ and the density of ν is $g_N/(1+\bar{r}^N)$. However, by the subadditivity of squared Wasserstein metric under rescaled convolution [22, 23], we can approximate

$W_2^2(\mu_N, \nu)$ by $W_2^2(f_N, g_N)$:

$$\begin{aligned} W_2^2(\mu_N, \nu) &\leq \left(\frac{1}{1+\bar{r}^N}\right)^2 W_2^2(f_N, g_N) \\ &\leq \left(\frac{1}{1-c}\right)^2 W_2^2(f_N, g_N) \\ &\leq \left(\frac{1}{1-c}\right)^2 (1+\bar{r}^N)^2 W_2^2(\mu_N, \nu) \\ &\leq \left(\frac{1+c}{1-c}\right)^2 W_2^2(\mu_N, \nu) \end{aligned}$$

and we conclude that $\mathbb{E}W_2^2(\mu_N, \nu)$ is $\mathcal{O}(\frac{1}{N})$. □

Remark 4. The L^2 norm is significantly more sensitive to noise in this setting since $\mathbb{E}L_2^2(f_N, g_N) = \mathbb{E}||f_N - g_N||_2^2 = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^N |r_i|^2\right) = O(1)$.

3.2. **Higher dimensions.** The analysis of the Wasserstein metric becomes much more difficult in higher dimensions. However, we can still analyze the effects of noise through the computation of an upper bound on the metric.

From the definition of the quadratic Wasserstein metric (1), it is clear that any mass-preserving map T satisfies the inequality

$$W_2^2(f, g) \leq \int |x - T(x)|^2 f(x) dx.$$

Consider the following two-dimensional example on the domain $\Omega = [0,1] \times [0,1]$ with the constant density function $g = 1$. Consider the noise function r such that for each $(x, y) \in \Omega$, $r(x, y)$ is a random variable with uniform distribution on $[-c, c]$, $0 < c < 1$. We define the noisy density function $f = g + r$ and assume that $\int_\Omega f = \int_\Omega g$.

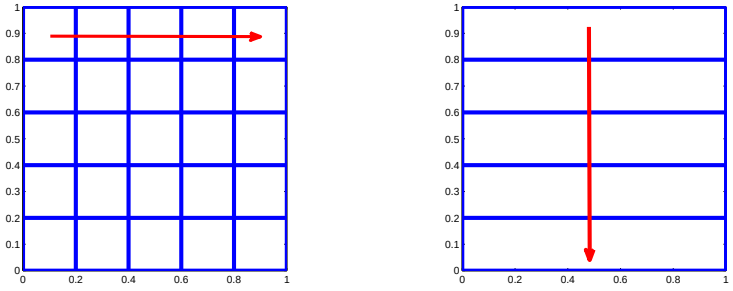
We use the Wasserstein metric to measure the difference between g and its noisy version f. Since strong convergence in measure implies convergence of the Wasserstein metric, we can approximate the density function f by the piecewise constant function f_N for the convenience of calculation.

$$f_N(x, y) = 1 + r_{ij}, \quad x_i = \frac{i}{N} < x \leq \frac{i+1}{N} = x_{i+1}, \quad y_j = \frac{j}{N} < y \leq \frac{j+1}{N} = y_{j+1}.$$

One approach to rearranging all the mass from f_N to g is to define T in two steps as in Figure 4. First, with y fixed, one can find the optimal map T_x that averages each row. This is equivalent to a 1D optimal transport problem. Each row i is mapped into a uniform density after rearrangement by the optimal map T_i. Secondly, with x fixed, one can average the the density values of all the rows. Again, this is a 1D optimal transport problem and we have an explicit form for the optimal map T_y. The resulting map T_N that rearranges f_N to g is $T_y \circ T_x$. Here $T_x = T_i$ for $x_{i-1} < x \leq x_i$, $i = 1, \ldots, N$.

The rearrangement determined by T_N is not optimal, but does provide an upper bound on the value of the Wasserstein metric. As in the 1D case, we can verify that

$$\mathbb{E}\left(\int |x - T_N(x)|^2 f_N(x) dx\right) = \mathcal{O}\left(\frac{1}{N}\right)$$



(a) (b)

FIGURE 4. (a) The optimal map for each row: $T_x = T_i$ for $x_i < x \leq x_{i+1}$ and (b) the optimal map in y direction: T_y

since $T_N = T_y \circ T_x$ is a composition of two optimal 1D maps. This leads to the upper bound

$$\begin{aligned}\mathbb{E}W_2^2(f_N, g) &\leq \mathbb{E}\left(\int_{\mathbb{R}^2} |\mathbf{x} - T_N(\mathbf{x})|^2 f_N(\mathbf{x}) d\mathbf{x}\right) \\ &= \mathbb{E}\left(\int_{\mathbb{R}}\int_{\mathbb{R}} |x - T_x(x)|^2 f_N(\mathbf{x}) dxdy\right) + \mathbb{E}\left(\int_{\mathbb{R}}\int_{\mathbb{R}} |y - T_y(y)|^2 f_N(\mathbf{x}) dydx\right) \\ &\propto \mathcal{O}\left(\frac{1}{N}\right).\end{aligned}$$

Finally, by the Lebesgue dominated convergence theorem,

$$\mathbb{E}W_2^2(f, g) \leq \lim_{N\to\infty} \mathbb{E}\left(\int_{\mathbb{R}^2} |\mathbf{x} - T_N(\mathbf{x})|^2 f_N(\mathbf{x}) d\mathbf{x}\right) = 0. \tag{17}$$

For higher dimension $n \geq 3$, we can similarly reduce the problem to several 1D optimal transport in terms of the dimensions. The ultimate goal is to find a particular map that is not optimal, but that provides an upper bound that goes to zero as the mesh is refined.

4. Numerical Computation of the Wasserstein metric

We are interested in computing the Wasserstein metric between two distributions f, g, which are supported on a rectangle X. This can be accomplished via the solution of the Monge-Ampère equation

$$\begin{cases} \det(D^2 u(x)) = f(x)/g(\nabla u(x)) + \langle u \rangle, & x \in X \\ \nabla u(x) \cdot \nu = x \cdot \nu, & x \in \partial X \\ u \text{ is convex.} \end{cases} \tag{18}$$

Remark 5. The Neumann boundary condition is easily generalised to the situation where f and g are supported on different rectangles [7].

The squared Wasserstein metric is then given by

$$W_2^2(f, g) = \int_X f(x) |x - \nabla u(x)|^2 \, dx. \tag{19}$$

We solve the Monge-Ampère equation numerically using an almost-monotone finite difference method relying on the following reformulation of the Monge-Ampère operator, which automatically enforces the convexity constraint [7].

$$\det{}^{+}(D^2u) = \min_{\{v_1,v_2\}\in V} \{\max\{u_{v_1,v_1},0\}\max\{u_{v_2,v_2},0\} + \min\{u_{v_1,v_1},0\} + \min\{u_{v_2,v_2},0\}\} \tag{20}$$

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

Equation (20) can be discretised by computing the minimum over finitely many directions $\{\nu_1,\nu_2\}$, which may require the use of a wide stencil. For simplicity and brevity, we describe a compact version of the scheme and refer to [7, 8] for complete details.

We begin by introducing the finite difference operators

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

In the compact version of the scheme, the minimum in (20) is approximated using only two possible values. The first uses directions aligning with the grid axes.

$$\begin{aligned}
MA_1[u] = \max\left\{\mathcal{D}_{x_1x_1}u,\delta\right\}\max\left\{\mathcal{D}_{x_2x_2}u,\delta\right\}\\
+ \min\left\{\mathcal{D}_{x_1x_1}u,\delta\right\} + \min\left\{\mathcal{D}_{x_2x_2}u,\delta\right\} - f/g\left(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u\right) - u_0.
\end{aligned} \tag{21}$$

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

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

$$\begin{aligned}
MA_2[u] = \max\left\{\mathcal{D}_{vv}u,\delta\right\}\max\left\{\mathcal{D}_{v^\perp v^\perp}u,\delta\right\} + \min\left\{\mathcal{D}_{vv}u,\delta\right\} + \min\left\{\mathcal{D}_{v^\perp v^\perp}u,\delta\right\}\\
- f/g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_{v}u + \mathcal{D}_{v^\perp}u), \frac{1}{\sqrt{2}}(\mathcal{D}_{v}u - \mathcal{D}_{v^\perp}u)\right) - u_0.
\end{aligned} \tag{22}$$

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

$$M_M[u] \equiv -\min\{MA_1[u], MA_2[u]\} = 0. \tag{23}$$

We also define a second-order non-monotone approximation, obtained from a standard centred difference discretisation,

$$M_N[u] \equiv -\left((\mathcal{D}_{x_1x_1}u)(\mathcal{D}_{x_2x_2}u) - (\mathcal{D}_{x_1x_2}u^2)\right) + f/g\left(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u\right) + u_0 = 0. \tag{24}$$

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

$$M_F[u] \equiv M_M[u] + \epsilon S\left(\frac{M_N[u] - M_M[u]}{\epsilon}\right) 0 \tag{25}$$

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

$$S(x) = \begin{cases} x & |x| \leq 1 \\ 0 & |x| \geq 2 \\ -x+2 & 1 \leq x \leq 2 \\ -x-2 & -2 \leq x \leq -1. \end{cases} \tag{26}$$

The Neumann boundary condition is implemented using standard one-sided differences.

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

$$W_2^2(f,g) \approx \sum_{j=1}^{m} (x_j - D_{x_j}u_h)^T \mathrm{diag}(f)(x_j - D_{x_j}u_h). \tag{27}$$

The computation of the discrete solution of (25) requires the solution of a large system of nonlinear algebraic equations. This is accomplished using Newton's method, which requires the Jacobian of the discrete scheme. The Jacobian of the filtered scheme can be expressed as

$$\nabla M_F[u] = \left(1 - S'\left(\frac{M_N[u] - M_M[u]}{\epsilon}\right)\right)\nabla M_M[u] + S'\left(\frac{M_N[u] - M_M[u]}{\epsilon}\right)\nabla M_N[u]. \tag{28}$$

The (formal) Jacobians of the monotone and non-monotone components are given by

$$\begin{aligned}
\nabla_u M_1[u] &= \left(\max\{\mathcal{D}_{x_2x_2},\delta\}\mathbb{1}_{\mathcal{D}_{x_1x_1}>\delta}+\mathbb{1}_{\mathcal{D}_{x_1x_1}\le\delta}\right)\mathcal{D}_{x_1x_1}\\
&+\left(\max\{\mathcal{D}_{x_1x_1},\delta\}\mathbb{1}_{\mathcal{D}_{x_2x_2}>\delta}+\mathbb{1}_{\mathcal{D}_{x_2x_2}\le\delta}\right)\mathcal{D}_{x_2x_2}\\
&-\frac{f}{g\left(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u\right)^2}\nabla g\left(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u\right)\cdot\left(\mathcal{D}_{x_1},\mathcal{D}_{x_2}\right)-\mathbb{1}_{x=x_0},\\
\nabla_u M_2[u] &= \left(\max\{\mathcal{D}_{v^\perp v^\perp},\delta\}\mathbb{1}_{\mathcal{D}_{vv}>\delta}+\mathbb{1}_{\mathcal{D}_{vv}\le\delta}\right)\mathcal{D}_{vv}\\
&+\left(\max\{\mathcal{D}_{vv},\delta\}\mathbb{1}_{\mathcal{D}_{v^\perp v^\perp}>\delta}+\mathbb{1}_{\mathcal{D}_{v^\perp v^\perp}\le\delta}\right)\mathcal{D}_{v^\perp v^\perp}\\
&-f/g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v u+\mathcal{D}_{v^\perp}u),\frac{1}{\sqrt{2}}(\mathcal{D}_v u-\mathcal{D}_{v^\perp}u)\right)^2\\
&\nabla g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v u+\mathcal{D}_{v^\perp}u),\frac{1}{\sqrt{2}}(\mathcal{D}_v u-\mathcal{D}_{v^\perp}u)\right)\\
&\cdot\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v+\mathcal{D}_{v^\perp}),\frac{1}{\sqrt{2}}(\mathcal{D}_v-\mathcal{D}_{v^\perp})\right)-\mathbb{1}_{x=x_0},\\
\nabla_u M_M[u] &= -\mathbb{1}_{M_M[u]=-M_1[u]}\nabla_u M_1[u]-\mathbb{1}_{M_M[u]=-M_2[u]}\nabla_u M_2 u],\\
\nabla_u M_N[u] &= -(\mathcal{D}_{x_2x_2}u)\mathcal{D}_{x_1x_1}-(\mathcal{D}_{x_1x_1}u)\mathcal{D}_{x_2x_2}+2(\mathcal{D}_{x_1x_2}u)\mathcal{D}_{x_1x_2}\\
&+\frac{f}{g\left(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u\right)^2}\nabla g\left(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u\right)\cdot\left(\mathcal{D}_{x_1},\mathcal{D}_{x_2}\right)+\mathbb{1}_{x=x_0}.
\end{aligned}$$

5. Computation of Frechet Gradient

Our goal is to minimise the Wasserstein metric between computed data $f(v)$ and observed data g, where f depends on a set of parameters v. In order to do this efficiently, we will require the gradient of the squared Wasserstein metric with respect to the unknown parameters.

Our main focus here is computation of the Fréchet gradient of the squared Wasserstein metric with respect to the data f, which is new in the context of full waveform inversion. The gradient needed for the minimization is then obtained through the composition

$$\nabla_f W_2^2 \nabla_v f.$$

As long as $\nabla_f W_2^2$ can be computed efficiently, techniques such as the adjoint state method can be used to efficiently construct the required gradient [16].

In the present work, our focus is on the use of optimal transportation techniques, rather than on the use of a particular forward model for producing the data $f(v)$. In the computations of section 6, we will present the minimisation for problems involving several different models. For simplicity, and to keep the focus on the properties of the Wasserstein metric, we will simply use a forward difference approximation to estimate $\nabla_v f$.

5.1. Linearisation of continuous problem.

Let $f : X \to (0,\infty)$, $g : X \to (0,\infty)$ be positive density functions defined on a rectangle X. We consider the squared Wasserstein metric $W_2(f,g)^2$ as a function of f,

$$d(f) = \int_X f(x)\left|x-\nabla u\right|^2 dx. \tag{29}$$

Here the function u satisfies the second boundary value problem for the Monge-Ampère equation

$$\begin{cases} \det(D^2u(x)) = f(x)/g(\nabla u(x)), & x \in X \\ \nabla u \cdot \nu = x \cdot \nu, & x \in \partial X \\ u \text{ is convex.} \end{cases} \tag{30}$$

Now we perturb f by an amount δf and investigate the resulting change δd in the function d. That is

$$d + \delta d = \int_X (f + \delta f) \left| x - \nabla(u + \delta u) \right|^2 \, dx$$

where $u + \delta u$ satisfies the Monge-Ampère equation with f replaced by $f + \delta f$.

Viscosity solutions of the Monge-Ampère equation are stable so we expect $\delta u = \mathcal{O}(\delta f)$.

The Monge-Ampère equation can be rewritten as

$$f + \delta f = g(\nabla(u + \delta u)) \det(D^2(u + \delta u)).$$

Linearising these terms, we have to first order

$$\begin{aligned} f + \delta f &= \left[g(\nabla u) + \nabla g(\nabla u) \cdot \nabla \delta u \right] \left[\det(D^2 u) + \operatorname{tr}((D^2 u)_{adj} D^2 \delta u) \right] \\ &= g(\nabla u) \det(D^2 u) + g(\nabla u) \operatorname{tr}((D^2 u)_{adj} D^2 \delta u) + \det(D^2 u) \nabla g(\nabla u) \cdot \nabla \delta u. \end{aligned}$$

Here we use the notation $A_{adj} = \det(A) A^{-1}$ to denote the adjugate of the matrix A.

Thus to first order, the variation in u satisfies the linear elliptic PDE

$$\mathcal{L}[\delta u] \equiv g(\nabla u) \operatorname{tr}((D^2 u)_{adj} D^2 \delta u) + \det(D^2 u) \nabla g(\nabla u) \cdot \nabla \delta u = \delta f. \tag{31}$$

Note that both u and $u + \delta u$ satisfy the boundary condition in (2). If the domain and target X are rectangles, this boundary condition can be rewritten as

$$\nabla u \cdot \nu = x \cdot \nu, \quad \nabla(u + \delta u) \cdot \nu = x \cdot \nu, \quad x \in \partial X$$

where ν is the unit outward normal to the domain. Thus the variation δu satisfies the homogeneous Neumann boundary condition

$$\nabla(\delta u) \cdot \nu = 0, \quad x \in \partial X. \tag{32}$$

Remark 6. There is an issue of existence and uniqueness for δu. It is also true that we need to restrict to mean zero perturbations δf in order for the Wasserstein metric to be defined. A term involving the average $\langle u \rangle$ can be added to the Monge-Ampère equation to correct the issues of existence/uniqueness.

To first order, the perturbed (squared) Wasserstein metric satisfies

$$\begin{aligned} d + \delta d &= \int_X (f + \delta f) \left| x - \nabla(u + \delta u) \right|^2 \, dx \\ &= \int_X \left[f \left| x - \nabla u \right|^2 + \delta f \left| x - \nabla u \right|^2 - 2f(x - \nabla u) \cdot \nabla \delta u \right] \, dx. \end{aligned}$$

Thus the first variation is given by

$$\delta d = \int_X \left[\left| x - \nabla u \right|^2 - 2f(x - \nabla u) \cdot \nabla \mathcal{L}^{-1} \right] \delta f$$

and the gradient of the squared Wasserstein metric is

$$\nabla d = \left| x - \nabla u_f \right|^2 - 2(\mathcal{L}^{-1})^*(\nabla \cdot (f(x - \nabla u_f))).$$

Computation of the gradient requires the solution of a single linear elliptic PDE.

5.2. **Linearisation of discrete problem.** As an alternative to the above, we consider the linearisation of the discretised version of the Wasserstein metric. Using the finite difference matrices introduced in section 4, we can express the discrete Wasserstein metric as

$$d(f) = \sum_{j=1}^{n} (x_j - D_{x_j} u_f)^T \mathrm{diag}(f)(x_j - D_{x_j} u_f) \tag{33}$$

where the potential u_f satisfies the discrete Monge-Ampère equation

$$\mathrm{M}[u_f] = 0.$$

The first variation of the squared Wasserstein metric as

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

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

$$\nabla M_F[u_f]\delta u = \delta f.$$

Here ∇M_F is the (formal) Jacobian of the discrete Monge-Ampère equation, which is already being inverted in the process of solving the Monge-Ampère equation via Newton's method (28). Then the gradient of the discrete squared Wasserstein metric can be expressed as

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

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

6. Computational Results

In this section, we provide examples of the minimization of the Wasserstein metric between given data g and a modeled signal $f(v)$ that depends on the unknown parameters v. Minimisation is performed using the Matlab function fmincon, equipped with the gradient described in subsection 5.2.

6.1. **Forward problem.** The 2-D acoustic wave equation in the time domain can be written as:

$$\begin{cases} \frac{1}{v(x,z)^2}\frac{\partial^2 u}{\partial t^2} - \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial z^2}\right) = 0, \\ u(x,z,t) = u_0(x,z), \quad t = 0, \\ u_t(x,z,t) = 0, \quad t = 0. \end{cases} \tag{34}$$

Here $u(x,z,t)$ is the wave field, $v(x,z)$ is the velocity, $u_0(x,z)$ is the initial wave field generated by Ricker wavelet signal[24].

We will use the velocity model to simulate a seismic survey. The wave equation (34) is solved by using finite difference scheme for a defined initial wave field.

(35)
$$u_{n,m}^{l+1} = -u_{n,m}^{l-1} + 2u_{n,m}^{l} + v_{n,m}^2 \Delta t^2 \left(\frac{u_{n+1,m}^{l} - 2u_{n,m}^{l} + u_{n-1,m}^{l}}{\Delta x^2} + \frac{u_{n,m+1}^{l} - 2u_{n,m}^{l} + u_{n,m-1}^{l}}{\Delta z^2} \right)$$

with the initial conditions

$$u_{n,m}^{-1} = f(n\Delta x, m\Delta z), \quad u_{n,m}^{0} = f(n\Delta x, m\Delta z). \tag{36}$$

Here $u_{n,m}^{l}$ is the wave field at the time $l\Delta t$ and at the spatial position $(n\Delta x, m\Delta z)$. $v_{n,m}$ is the velocity at $(n\Delta x, m\Delta z)$. The step size Δt is chosen to satisfy the numerical stability condition:

$$\min(\Delta x, \Delta z) > \sqrt{2}\Delta t \max(v). \tag{37}$$

The wave field $u(x, 0, t)$ recovered at the surface forms the two-dimensional seismic data.

We cannot directly compute the Wasserstein metric between two such wave fields since these are not probability measures. Some additional processing is needed to ensure that the data $f(x, t)$ is strictly positive and has total mass one. Instead, we work with something akin to a local amplitude by defining

$$\tilde{f}(x, t) = \sqrt{\int_{t-\epsilon}^{t+\epsilon} u(x, 0, s)^2 \, ds}$$

where $\epsilon = 10\Delta t$. Finally, this profile is normalised to produce a density function $f(x, t)$ that has unit mass.

6.2. **Single layer model.** We first consider a material composed of a single layer of depth d and velocity v. We define the data $f_{d,v}(s, t)$ to be the resulting data, which we obtain by solving the wave equation for $u_{d,v}$ and processing the results.

We consider the particular case of $d = 2$, $v = 1$. In order to define the target profile g, which mimics the observed data, we add noise $N(s, t)$ chosen uniformly at random from $[-M, M]$,

$$\tilde{g}(s, t) = \max\{u_{2,1}(s, t) + N(s, t), 0\},$$

where M is approximately 2% the maximum value of $f_{2,1}$. See Figure 6.2. Then our goal is to determine d and v that minimise

$$W_2^2(f_{d,v}, g).$$

We initialise with the guess $d = 2.5$ and $v = 1.75$ and perform minimisation over the parameters d and v^{-1}. The convergence history is displayed in Figure 8. Despite the noise in the target profile, we recover the parameters $d = 2.2157$ and $v = 1.0953$ after fifteen iterations, with a squared Wasserstein metric of 3.36×10^{-4}. (The required stepsize in the minimization algorithm became too small to improve appreciably beyond this). For reference, comparison of the noisy target with the exact signal $f_{2,1}$ (without noise) yielded a squared Wasserstein metric of 7.49×10^{-4}, so that the error in the recovered parameters can be explained by the noise.

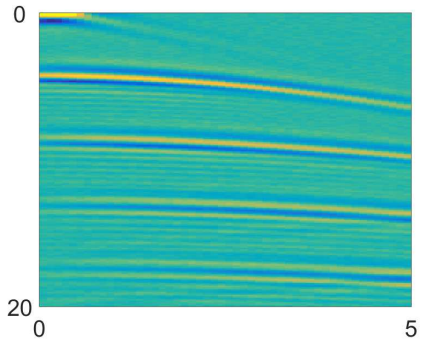
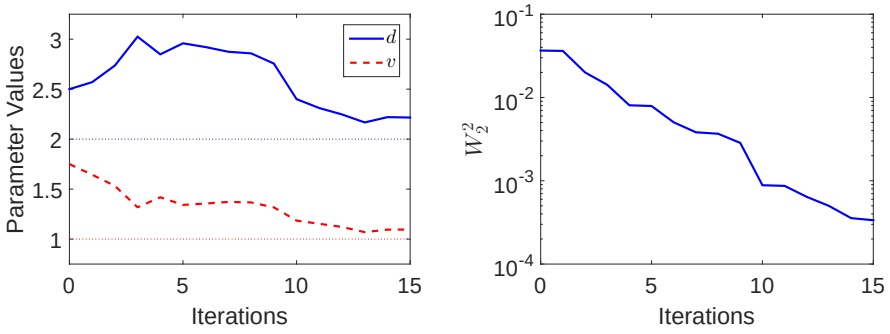


FIGURE 5. A signal produced from a single layer model with added noise.



(a) (b)

FIGURE 6. Convergence history for a single layer model.

6.3. **Two layer model.** Next, we consider the case where the material is composed of two different layers. The top layer has depth d_1 and velocity v_1 while the bottom layer has depth d_2 and velocity v_2; see Figure 7(a).

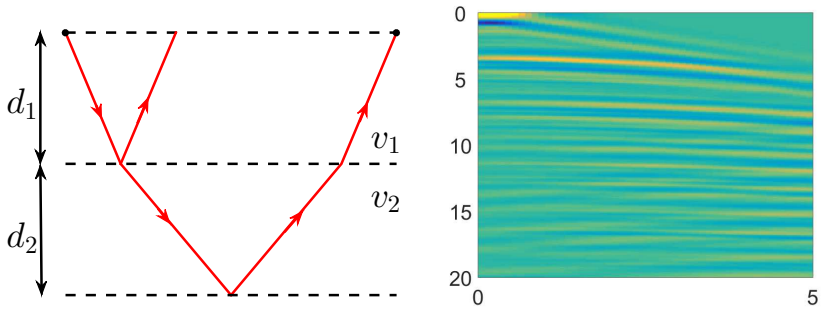
We look at the particular case where the given target density g is defined by the parameter values

$$d_1 = 0.75, \quad v_1 = 1, \quad d_2 = 1, \quad v_2 = 1.5.$$

As in the previous example, we add noise to this target. The resulting signal is shown in Figure 7(b).

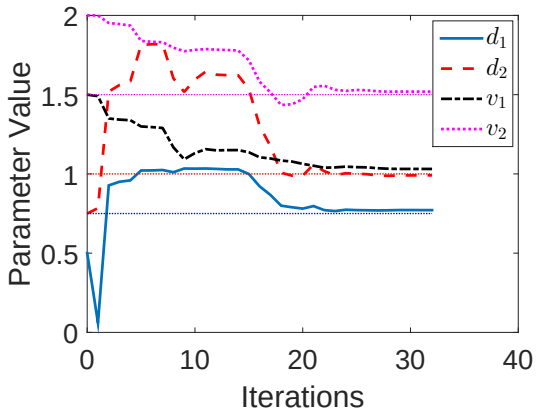
In this case, the distance W_2 depends on the four parameters $d_1, v_1^{-1}, d_2, v_2^{-1}$. We initialise with the guess $d_1 = 0.5$, $v_1 = 1.5$, $d_2 = 0.75$, and $v_2 = 2$. After 33 iterations, we recover the parameter values $d_1 = 0.772$, $d_2 = 0.991$, $v_1 = 1.0318$, and $v_2 = 1.519$ with a squared misfit value of 2.06×10^{-5}. The convergence history is presented in Figure 8.

As noted in [6], when the model involves both depth and velocity, the resulting distance can contain narrow valleys, and computing the minimum can require small

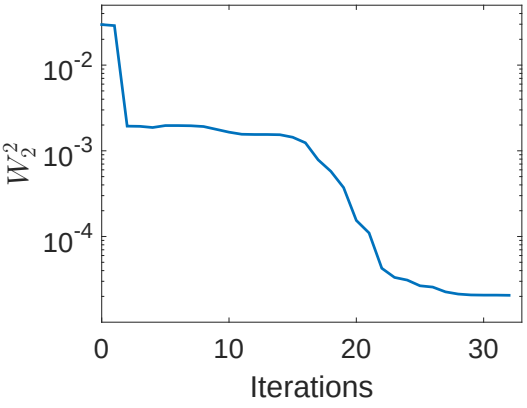


(a) (b)

FIGURE 7. (a) A two-layer material and (b) the resulting signal.



(a)



(b)

FIGURE 8. Convergence history for a two-layer material.

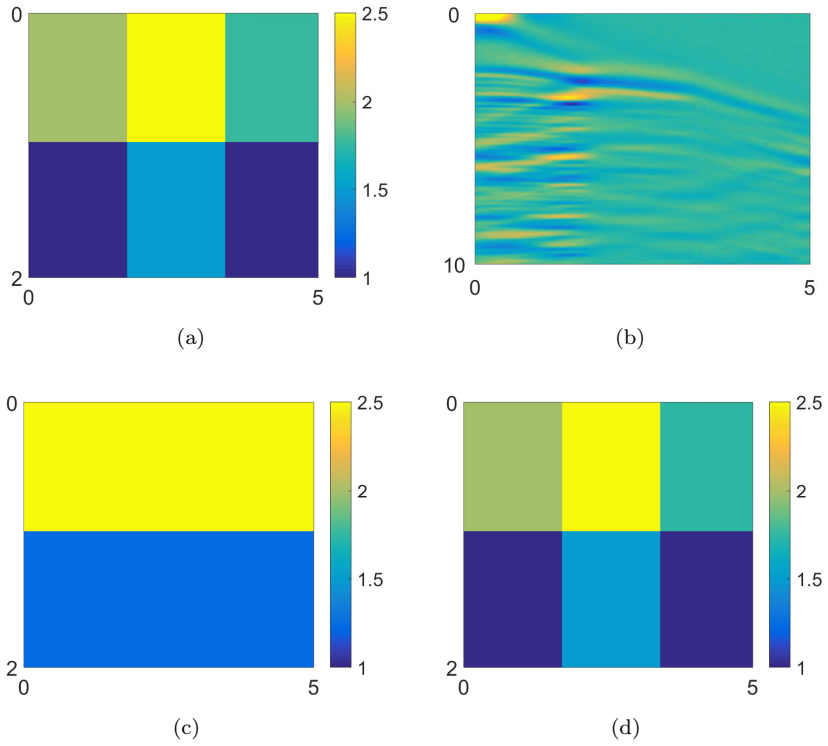


FIGURE 9. (a) A six parameter velocity model used to generate (b) a target signal g. (c) Initial and (d) computed velocity.

stepsizes. We were still able to effectively compute the minimum in this setting, but we expect that quasi-Newton methods would enable even faster convergence.

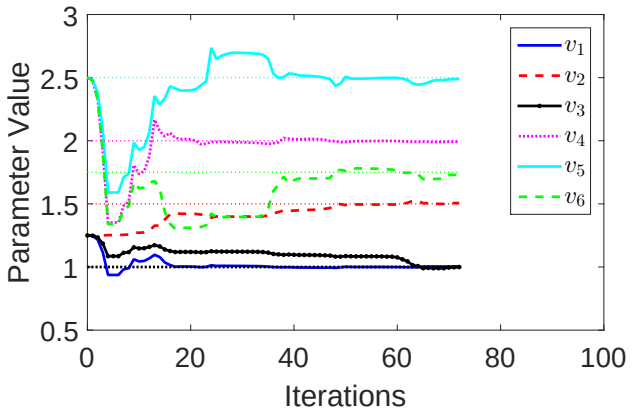
6.4. **Six Parameter model.** We next consider the case of a piecewise constant material. See Figure 9 for the set-up.

We look at the particular case where the given target density g is defined by the parameter values

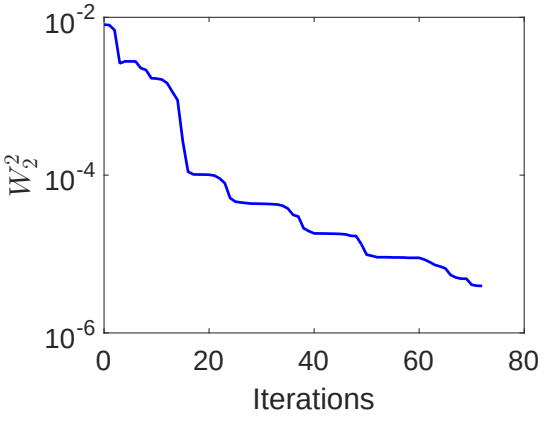
$$v_1 = 1, \quad v_2 = 1.5, \quad v_3 = 1, \quad v_4 = 2, \quad v_5 = 2.5, \quad v_6 = 1.75.$$

As in the previous example, we add noise to this target. The resulting signal is shown in Figure 9(b).

In this case, the distance W_2 depends on the six parameters $1/v_i$, $i = 1, \ldots, 6$. We initialise with the guess $v_1 = v_2 = v_3 = 1.25$ and $v_4 = v_5 = v_6 = 2.5$. In this example, which depends only on velocity and not on depth, the convergence proceeds without the need for very small stepsizes that we observed in the previous example. After 72 iterations, we recover the parameter values $v_1 = 1.0034$, $v_2 = 1.5058$, $v_3 = 0.9996$, $v_4 = 1.9932$, $v_5 = 2.4889$, and $v_6 = 1.7296$ with a squared misfit value of 3.94×10^{-6}. The convergence history is presented in Figure 10.



(a)



(b)

FIGURE 10. Convergence history for a six parameter model.

For reference, comparison of the noisy target with the exact signal (without noise) yielded a squared Wasserstein metric of 4.82×10^{-6}.

6.5. **Twelve Parameter model.** We again consider a piecewise constant velocity model, but this time increase the number of parameters to twelve. See Figure 11 for the set-up used to construct the (noisy) target density g, as well as the resulting signal.

In this case, the distance W_2 depends on the twelve parameters $1/v_i$, $i = 1, \ldots, 12$. We initialise with the guess $v = v_g + 0.25$. After 132 iterations, we recover the twelve parameters with a maximum error of $\|v - v_g\|_\infty = 0.0091$ and a squared misfit value of 2.10×10^{-6}. For reference, comparison of the noisy target with the exact signal (without noise) yielded a squared Wasserstein metric of 3.16×10^{-6}, which suggests that the error in the recovered parameter values is due to noise in the data. The convergence history is presented in Figure 12. The simple

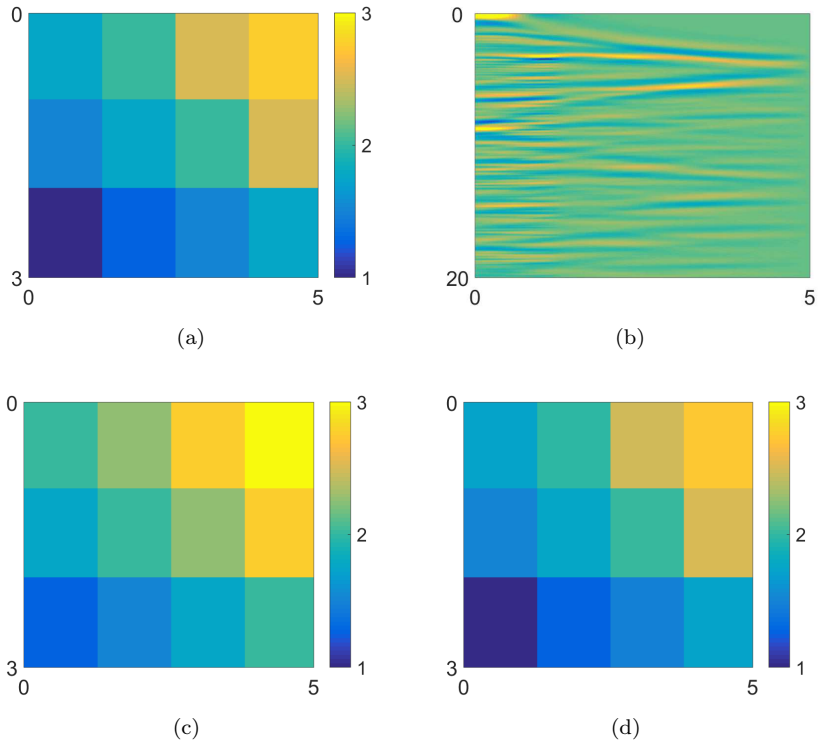


FIGURE 11. (a) A twelve parameter velocity model used to generate (b) a target signal g. (c) Initial and (d) computed velocity.

models in subsection 6.2 and subsection 6.4 were included to indicate how the result depend on model complexity.

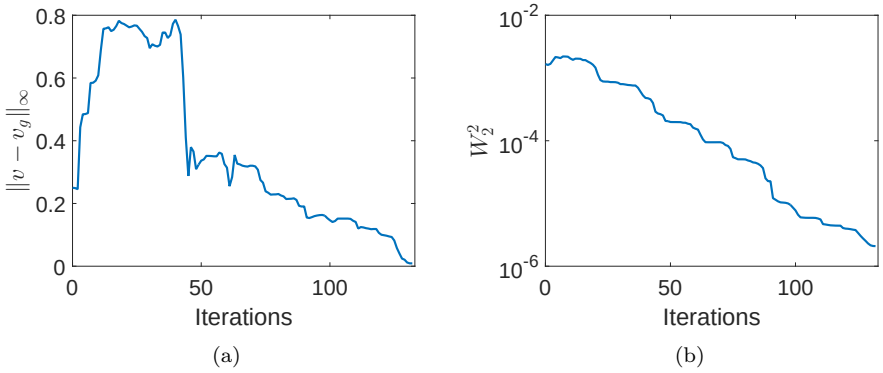


FIGURE 12. Convergence history for a twelve parameter model.

7. Conclusions

In this paper, we demonstrate several advantages of the Wasserstein metric as a measure of misfit between seismic signals in connection to full waveform inversion. In particular, we proved that this distance is convex with respect to several common transformations and is less sensitive to noise than the L_2 distance. Additionally, the Frechét gradient is easily computed, which makes the Wasserstein metric extremely promising for optimization and thus for seismic inversion problems. Simple numerical examples demonstrate the efficiency of using this metric.

A natural direction for future research is increasing the efficiency of the computation with quasi Newton techniques and parallelization, and thus being able to apply the method on more realistic seismic applications.

References

[1] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[2] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. 2013. Submitted.

[3] D. P. Bertsekas. Auction algorithms for network flow problems: a tutorial introduction. *Comput. Optim. Appl.*, 1(1):7–66, 1992.

[4] D. Bosc. Numerical approximation of optimal transport maps. *SSRN*, 2010.

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

[6] B. Engquist and B. D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5), 2014.

[7] B. D. Froese. A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions. *SIAM J. Sci. Comput.*, 34(3):A1432–A1459, 2012.

[8] B. D. Froese and A. M. Oberman. Convergent filtered schemes for the Monge- Ampère partial differential equation. *SIAM J. Numer. Anal.*, 51(1):423–444, 2013.

[9] V. Isakov. *Inverse problems for partial differential equations*, volume 127. Springer Science & Business Media, 2006.

[10] M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

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

[12] I. Masoni, R. Brossier, J. Virieux, and J. L. Boelle. Alternative misfit functions for FWI applied to surface waves. In *75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013*, 2013.

[13] R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–324, 1995.

[14] R. J. McCann. Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and geometry of metric measure spaces: lecture notes of the Séminaire de Mathématiques Supérieure (SMS) Montréal*, pages 145–180, 2011.

[15] W. Menke. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, 1984.

[16] R.-E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006.

[17] J. A. Scales, P. Docherty, and A. Gersztenkorn. Regularisation of nonlinear inverse problems: imaging the near-surface weathering layer. *Inverse Problems*, 6(1):115, 1990.

[18] P. Stefanov and G. Uhlmann. Recovery of a source term or a speed with one measurement and applications. *Transactions of the American Mathematical Society*, 365(11):5737–5758, 2013.

[19] J. Sylvester and G. Uhlmann. A global uniqueness theorem for an inverse boundary value problem. *Annals of mathematics*, pages 153–169, 1987.

[20] A. Tarantola. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266, 1984.

[21] A. Tarantola. Inverse problems theory. *Methods for Data Fitting and Model Parameter Estimation. Elsevier, Southampton*, 1987.
[22] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
[23] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[24] Wensheng Zhang and Jia Luo. Full-waveform velocity inversion based on the acoustic wave equation. *American Journal of Computational Mathematics*, 3(03):13, 2013.

Department of Mathematics and ICES, The University of Texas at Austin, 1 University Station C1200, Austin, TX 78712 USA
E-mail address: engquist@math.utexas.edu

Department of Mathematical Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102 USA
E-mail address: bdfroese@njit.edu

Department of Mathematics and ICES, The University of Texas at Austin, 1 University Station C1200, Austin, TX 78712 USA
E-mail address: yunanyang@math.utexas.edu