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 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 Fréchet gradient for use in the optimization. Numerical examples are given.

Key words. Optimal transport, Wasserstein metric, computational seismology, full waveform inversion.

AMS subject classifications. 35J96, 49N45, 65K10, 65N06, 86A15.

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 [12, 20]. This inverse method follows the common strategy of PDE constrained optimization. In this context, the goal is to recover the unknown wave velocity $v(x,z)$ from the resulting wavefield $u(x,z,t)$. In practice, measurments are available only at the surface, and the velocity field needs to be recovered from the surface measurement

$$g = u(x,0,t).$$

In two-dimensions, for example, this can be modeled by the acoustic wave equation in the time domain:

$$\begin{cases} u_{tt}(x,z,t) - v(x,z)^2 \left(u_{xx}(x,z,t) + u_{zz}(x,z,t) \right) = 0, \\ u(x,z,0) = u_0(x,z), \\ u_t(x,z,0) = 0, \end{cases} \tag{1.1}$$

where $u_0(x,z)$ is the initial wave field generated by a Ricker wavelet signal [24]. For a given wavefield v, the solution of this wave equation yields the simulated data

$$f(v) = u(x,0,t). \tag{1.2}$$

[†]Department of Mathematics and ICES, The University of Texas at Austin, 1 University Station C1200, Austin, TX 78712 USA (engquist@math.utexas.edu). https://www.ices.utexas.edu/people/107

[‡]Department of Mathematical Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102 USA (bdfroese@njit.edu). https://web.njit.edu/~bdfroese/

[§]Department of Mathematics, The University of Texas at Austin, University Station C1200, Austin, TX 78712 USA (yunanyang@math.utexas.edu). https://www.ma.utexas.edu/users/yunanyang/

In an ideal setting, the observed data also solves this forward problem so that

$$g = f(v^*),$$

where v^* is the true velocity field. However, this is unlikely in practice due to the presence of noise, measurement errors, and modeling errors. Instead, the goal of full waveform inversion is to estimate the the true velocity field through the solution of the optimization problem

$$\tilde{v} = \underset{v}{\operatorname{argmin}} \, d(f(v), g), \tag{1.3}$$

where $d(f, g)$ is some measure of the misfit between two signals.

Two primary concerns in full waveform inversion are the well-posedness of the underlying model recovery problem and the suitability of the misfit $d(f, g)$ for minimization. In this work we will focus on properties of the misfit measure d and not on the overall question of uniqueness and stability of the inverse problem. For additional details on uniqueness and stability, we refer to [10, 18, 19].

The L_2 norm is often used to measure the misfit, which typically generates many local minima and is thus unsuitable for minimization. This problem is exacerbated by the fact that measured signals usually suffer from noise in the measurements [13].

In [7], 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, p. 207]. The mathematical definition of the distance between the distributions $f : X \to \mathbb{R}^+$, $g : Y \to \mathbb{R}^+$ can be formulated as

$$W_2^2(f, g) = \inf_{T \in \mathcal{M}} \int_X |x - T(x)|^2 f(x) \, dx, \tag{1.4}$$

where $\mathcal{M}$ is the set of all maps that rearrange the distribution f into g.

We cannot directly compute the Wasserstein metric between two wave fields since these are not probability measures. Some additional processing is needed to ensure that the signals are strictly positive and have unit mass. This can be easily done by separately comparing the positive and negative parts of the signals, which are then rescaled to have mass one. We define $f^+ = \max\{f, 0\}$, $f^- = \max\{-f, 0\}$, and $\langle f \rangle = \int_X f(x) \, dx$. With this notation, we propose solving the optimization problem (1.3) using the misfit

$$d(f, g) = W_2^2 \left(\frac{f^+}{\langle f^+ \rangle}, \frac{g^+}{\langle g^+ \rangle} \right) + W_2^2 \left(\frac{f^-}{\langle f^- \rangle}, \frac{g^-}{\langle g^- \rangle} \right). \tag{1.5}$$

It is our goal to prove several desirable properties relating to convexity and insensitivity to noise, which were briefly discussed in [7]. 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 [7]. 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 [7] that illustrates the advantage of the Wasserstein metric. Consider the misfit between the simple wavelet f in Figure 1.1(a) and another wavelet shifted by a distance s. Figures 1.1(b)–1.1(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.
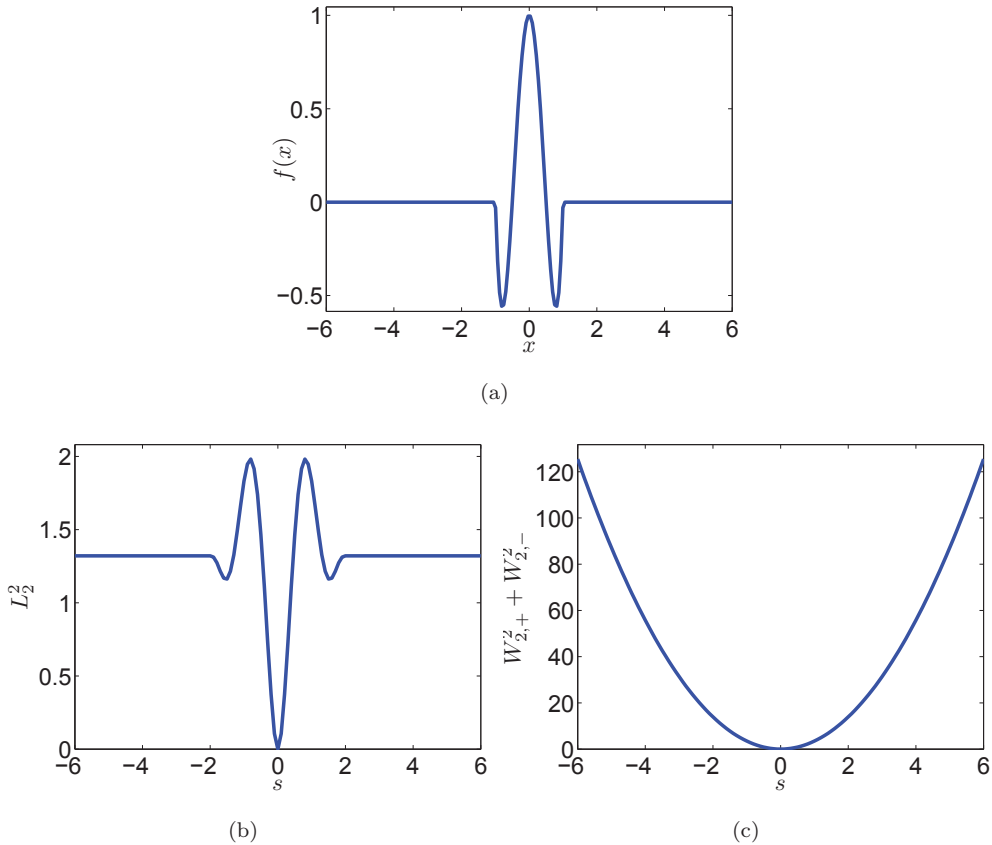

(a)

(b)

(c)

FIG. 1.1. *(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) $W_2^2(f^+,g^+) + W_2^2(f^-,g^-)$[7].*

Earlier algorithms for the numerical computation of the Wasserstein metric required a large number of operations [2, 4, 5]. The optimal transportation problem can be rigorously related to the following Monge–Ampère equation [6, 11], which enables the construction of more efficient methods for computing the Wasserstein metric.

$$\begin{cases} \det(D^2 u(x)) = f(x)/g(\nabla u(x)) + \langle u \rangle, & x \in X \\ u \text{ is convex.} \end{cases} \tag{1.6}$$

The Wasserstein metric is then given by

$$W_2^2(f,g) = \int_X |x - \nabla u(x)|^2 f(x)\, dx. \tag{1.7}$$

There are now fast and robust numerical algorithms for the solution of the system (1.6), and thus for the computation of W_2^2, and these form the basis for our numerical techniques [3].

In Section 2, we will study the convexity of the quadratic Wasserstein metric with respect to shift, dilation, and partial amplitude change. Errors between simulated and observed data in the form of shifts and dilations can occur naturally from an incorrect velocity model, while inaccurate measurements or variations in the strength of reflecting surfaces 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, p. 80].

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 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 describe a method for numerically obtaining the gradient of the Wasserstein metric by first discretizing the metric, then linearizing the result. This 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 any need to regularize the problem.

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 1.1(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 variations that are common in the context of seismic exploration. In particular, 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 shift and dilation are typical effects of variations in the velocity v, as can be seen in a simple example. A 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.

Local changes in amplitude can originate from variations in the strength of reflecting surfaces. A material composed of layers of different materials will yield a velocity field that is (approximately) piecewise constant. The strength of the reflection at a discontinuity is, in turn, related to the velocity field in each layer. An incorrect estimation of the velocity field will then lead to larger or smaller local amplitudes in the resulting wavefield.

We note that any shifts, dilations, and amplitude changes in a signal f will correspond to a similar transformation in the positive and negative parts of f. Thus in the results below, it is sufficient to assume that the original profile f is non-negative. The results below will still consider any changes in mass that result from the transformations.

2.1. Convexity with respect to shift. We begin by assuming looking at the effects of shifting a density function f, which has no effect on the total mass.

THEOREM 2.1 (Convexity of shift). *Suppose f and g are probability density functions of bounded second moment. Let T be the optimal map that rearranges f into g. If $f_s(x) = f(x - s\eta)$ for $\eta \in \mathbb{R}^n$, 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 the shift size s.*

The proof relies on the concept of cyclical monotonicity, which can be used to characterize an optimal map.

DEFINITION 2.1 (Cyclical monotonicity). *We say that a map $T : X \subset \mathbb{R}^n \to \mathbb{R}^n$ is cyclically monotone if for any $m \in \mathbb{N}^+$, $x_i \in X$, $1 \le i \le m$, $x_0 \equiv x_m$,*

$$\sum_{i=1}^m x_i \cdot T(x_i) \ge \sum_{i=1}^m x_i \cdot T(x_{i-1}). \tag{2.1}$$

THEOREM 2.2 (Optimality criterion for quadratic cost [1, Theorem 2.13]). *Let f and g be probability density functions supported on sets X and Y, respectively. A mass-preserving map $T : X \to Y$ minimizes the quadratic optimal transport cost if and only if it is cyclically monotone.*

Proof. **(Proof of Theorem 2.1.)** By construction, the map T_s rearranges f_s into g. The cyclical monotoncity of T_s follows immediately from the cyclical monotoncity of T

$$\sum_{i=1}^m x_i \cdot T_s(x_i) = \sum_{i=1}^m [(x_i - s\eta) + s\eta] \cdot T(x_i - s\eta)$$
$$\ge \sum_{i=1}^m [(x_i - s\eta) + s\eta] \cdot T(x_{i-1} - s\eta)$$
$$= \sum_{i=1}^m x_i \cdot T_s(x_{i-1}).$$

Then the squared Wasserstein metric can be expressed as

$$W_2^2(f_s, g) = \int |x - T_s(x)|^2 f_s(x) dx = \int |x - T(x - s\eta)|^2 f(x - s\eta) dx$$
$$= W_2^2(f, g) + s^2 |\eta|^2 + 2s \int \eta \cdot (x - T(x)) f(x) dx. \tag{2.2}$$

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

2.2. Convexity with respect to dilation. Next we consider the convexity of the Wasserstein metric with respect to dilations or contractions of the density functions. We begin by characterizing the optimal map in this setting.

LEMMA 2.1 (Optimal map for dilation). *Assume f and g are probability density functions of bounded second moment satisfying $f(x) = \det(A) g(Ax)$, where A is a symmetric*

positive definite matrix. Then the optimal transport map rearranging $f(x)$ into $g(y)$ is $T(x) = Ax$.

Proof. Again, the cyclical monotonicity condition of Theorem 2.2 is the key to verifying optimality.

Since A is symmetric positive definite, it has a unique Cholesky decomposition $A = L^T L$ for some upper triangular matrix L. Then for any $x_i \in X$,

$$\sum_{i=1}^{m} x_i \cdot (T(x_i) - T(x_{i-1})) = \frac{1}{2} \sum_{i=1}^{m} (x_{i-1}^T L^T L x_{i-1} + x_i^T L^T L x_i - 2x_i^T L^T L x_{i-1})$$

$$= \frac{1}{2} \sum_{i=1}^{m} |Lx_i - Lx_{i-1}|^2$$

$$\geq 0,$$

which verifies the optimality condition.   □

REMARK 2.1. 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 satisfy the symmetry condition $g(x,y) = g(-x,-y)$. Then the optimal map from $f(x) = g(Ax)$ to g is the identity function instead of $T(x) = Ax$.

Convexity is a separate issue as it depends on the parameterization. A special case of dilation occurs when A is a diagonal matrix. The following theorem is a generalization where the dilation need not occur along coordinate directions.

THEOREM 2.3 (Convexity with respect to dilation). *Assume $f(x)$ is a probability density function and $g(y) = f(A^{-1}y)$ where A is a symmetric positive definite matrix. Then the squared Wasserstein metric $W_2^2(f, g/\langle g \rangle)$ is convex with respect to the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A.*

Proof. In order to define the Wasserstein metric, it is necessary to work with the normalized density $g/\langle g \rangle = \det(A)^{-1} f(A^{-1}y)$. By Lemma 2.1, the optimal mapping is

$$T(x) = Ax = O\Lambda O^T$$

where O is an orthogonal matrix and Λ is a diagonal matrix whose entries are the eigenvalues $\lambda_1, \ldots, \lambda_n$. Then the squared Wasserstein metric can be expressed as

$$W_2^2 \left(f, \frac{g}{\langle g \rangle} \right) = \int f(x) |x - Ax|^2 \, dx$$

$$= \int f(x) x^T O(I - \Lambda)^2 O^T x \, dx$$

$$= \int f(Oz) z^T (I - \Lambda)^2 z \, dz,$$

which is convex in $\lambda_1, \ldots, \lambda_n$.   □

REMARK 2.2. 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.   Finally, we consider the problem where a profile f is derived from g, but with a decreased amplitude in part of the domain. That is, we suppose that the domain is decomposed into $\Omega = \Omega_1 \cup \Omega_2$ with $\Omega_1 \cap \Omega_2 = \emptyset$. For an amplitude loss parameter $0 \leq \beta \leq 1$ we suppose that f depends on the probability density function g via

$$f_\beta(x) = \begin{cases} \beta g(x), & x \in \Omega_1 \\ g(x), & x \in \Omega_2. \end{cases} \tag{2.3}$$

THEOREM 2.4 (Convexity with respect to partial amplitude loss).   *The squared Wasserstein metric $W_2^2(f_\beta / \langle f_\beta \rangle, g)$ is a convex function of the parameter β.*

In order to prove this result, we introduce an alternative form of the rescaled density in terms of a parameter $-1 \leq \alpha \leq 0$.

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

where

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

Note that h_α is non-negative and has unit mass by construction, with $h_0 = g$.

The two different families of parameters are connected as follows.

LEMMA 2.2 (Parameterization of amplitude loss).   *Define the parameterization function*

$$\alpha(\beta) = \frac{\beta}{\beta \int_{\Omega_1} g + \int_{\Omega_2} g} - 1.$$

Then $\alpha : [0,1] \to [-1,0]$ is concave and the associated density functions are related through

$$\hat{f}_\beta \equiv \frac{f_\beta}{\langle f_\beta \rangle} = h_{\alpha(\beta)}.$$

The proof of convexity with respect to β will come via convexity with respect to α.

LEMMA 2.3 (Convexity with respect to partial amplitude change).   *The squared Wasserstein metric $W_2^2(h_\alpha, g)$ is a convex function of the parameter α.*

Proof.   Choose any $\alpha_1, \alpha_2 \in [-1,0]$ and $s \in [0,1]$. From convexity of the Monge–Kantorovich minimization problem [22, p. 220], we have

$$W_2^2(sh_{\alpha_1} + (1-s)h_{\alpha_2}, g) \leq sW_2^2(h_{\alpha_1}, g) + (1-s)W_2^2(h_{\alpha_2}, g). \tag{2.4}$$

We can calculate

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

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

$$= h_{s\alpha_1+(1-s)\alpha_2}.$$

Thus we can rewrite Equation (2.4) as

$$W_2^2(h_{s\alpha_1+(1-s)\alpha_2}, g) \le sW_2^2(h_{\alpha_1}) + (1-s)W_2^2(h, f_{\alpha_2}, g) \tag{2.5}$$

and the Wasserstein metric $W_2^2(h_\alpha, g)$ is convex with respect to α.   □

A simple consequence of this result is that the misfit is a decreasing function of α.

LEMMA 2.4 (Misfit is non-increasing).   Let $-1 \le \alpha_1 < \alpha_2 \le 0$.   Then $W_2^2(h_{\alpha_2}, g) \le W_2^2(h_{\alpha_1}, g)$.

Proof.   Define the parameter $s \in [0,1]$ by $s = \alpha_2/\alpha_1$. Then we can use the convexity result of Lemma 2.3 to compute

$$\begin{aligned}
W_2^2(h_{\alpha_2}, g) &= W_2^2(h_{s\alpha_1+(1-s)\cdot 0}, g) \\
&\le sW_2^2(h_{\alpha_1}, g) + (1-s)W_2^2(h_0, g) \\
&\le W_2^2(h_{\alpha_1}, g),
\end{aligned}$$

where we have used the fact that $h_0 = g$.   □

Using these lemmas, we can now establish convexity with respect to the natural amplitude loss parameter β.

Proof. **(Proof of Theorem 2.4.)** Choose any $\beta_1, \beta_2, s \in [0,1]$. From the concavity of $\alpha(\beta)$ we have

$$\alpha(s\beta_1 + (1-s)\beta_2) \ge s\alpha(\beta_1) + (1-s)\alpha(\beta_2).$$

Applying Lemmas 2.3–2.4 we can compute

$$\begin{aligned}
W_2^2(\hat{f}_{s\beta_1+(1-s)\beta_2}, g) &= W_2^2(h_{\alpha(s\beta_1+(1-s)\beta_2)}, g) \\
&\le W_2^2(h_{s\alpha(\beta_1)+(1-s)\alpha(\beta_2)}, g) \\
&\le sW_2^2(h_{\alpha(\beta_1)}, g) + (1-s)W_2^2(h_{\alpha(\beta_2)}, g) \\
&= sW_2^2(\hat{f}_{\beta_1}, g) + (1-s)W_2^2(\hat{f}_{\beta_2}, g),
\end{aligned}$$

which establishes the convexity.   □

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 misfit function. 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. A simple one-dimensional example given by Villani [22, Exercise 7.11] shows that $W_2^2(f_\epsilon, g) = O(\epsilon^2)$ for $f_\epsilon = \left(1 + \sin\frac{2\pi x}{\epsilon}\right)$ and $g = 1$ on $[0,1]$. A numerical example without analysis was given in [7].

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 3.2. Then it is well known [22, Theorem 2.18] that the optimal transportation cost is

$$W_2^2(f,g)=\int_0^1 |F^{-1}(t)-G^{-1}(t)|^2\,dt. \tag{3.1}$$

If additionally the target density g is positive, then the optimal map from f to g is given by

$$T(x)=G^{-1}(F(x)).$$
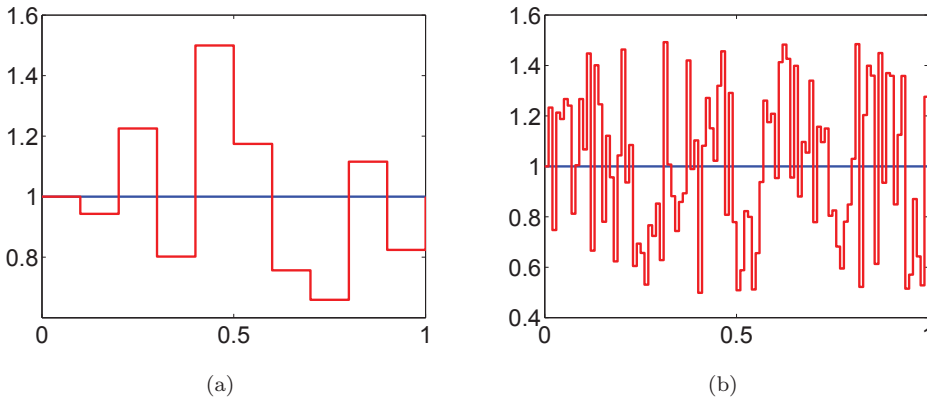
FIG. 3.1. *Densities g (blue) and f_N (red, $c=-0.5$) for (a) $N=10$ and (b) $N=100$.*

THEOREM 3.1 (Insensitivity to noise in 1D). *Let g be a positive probability density function on $[0,1]$ and choose $0<c<\min g$. Let $f_N(x)=g(x)+r^N(x)$, which contains piecewise constant additive noise r_N drawn from the uniform distribution $U[-c,c]$. Then $\mathbb{E}W_2^2(f_N/\langle f_N\rangle,g)=\mathcal{O}(\frac{1}{N})$.*

Without loss of generality, we take $g=1$ on $[0,1]$. Figure 3.1 shows the effect of the noise. For $x\in\left(\frac{i-1}{N},\frac{i}{N}\right]$, $r^N(x)\equiv r_i$, with each r_i drawn from $U[-c,c]$. As $N\to\infty$, $r^N(x)$ 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 $U[-c,c]$.

Proof. **(Proof of Theorem 3.1.)** For each i, r_i is a random variable of uniform distribution $U[-c,c]$, $0<c<\min g$. Thus, we have $\mathbb{E}r_i=0$ and $\mathbb{E}\bar r=0$.

Let $h=1/N$ and $x_i=ih$ for $i=0,\dots,N$. 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 the constant $g_N=1+\bar r^N$, which share the same mass.
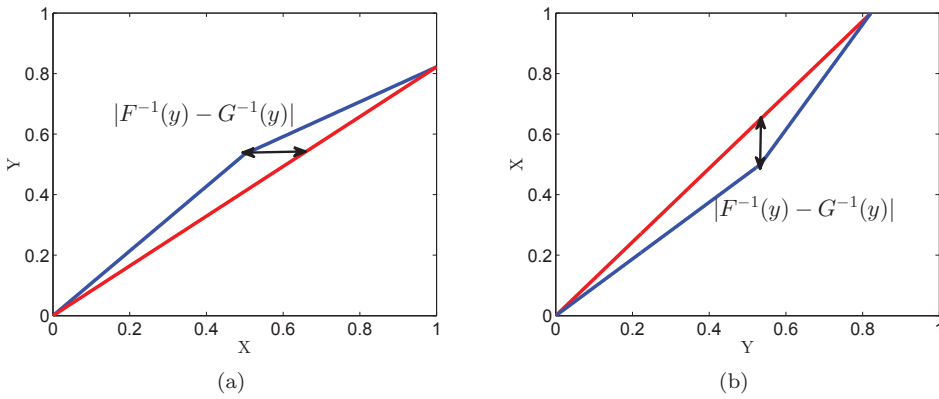
FIG. 3.2. *(a) Cumulative distribution functions $F(x)$ (red), $G(x)$ (blue) and (b) the inverse functions $F^{-1}(y)$ (red), $G^{-1}(y)$ (blue).*

In order to make use of Equation (3.1), we derive the cumulative distribution function and its inverse for both f_N and g_N:

$$F_N(x) = \sum_{j=1}^{i-1}(1+r_j)h + (1+r_i)(x-x_{i-1}), \qquad\qquad x \in (x_{i-1}, x_i]$$

$$G_N(x) = (1+\bar{r}^N)x, \qquad\qquad x \in [0,1]$$

$$F_N^{-1}(x) = \frac{x + \left((i-1)r_i - \sum_{j=1}^{i-1}r_j\right)h}{1+r_i}, \qquad x \in \left(\sum_{j=1}^{i-1}(1+r_j)h, \sum_{j=1}^{i}(1+r_j)h\right]$$

$$G_N^{-1}(x) = \frac{x}{1+\bar{r}^N}, \qquad\qquad x \in [0, 1+\bar{r}^N],$$

where $1 \leq i \leq N$.

Then we can bound the squared Wasserstein metric by

$$W_2^2(f_N, g_N) = \int_0^{1+\bar{r}} |F_N^{-1}(t) - G_N^{-1}(t)|^2 dt \leq \frac{2h^3}{(1-c)^2} \sum_{i=1}^{N}\left(\sum_{l=1}^{i}r_l - ih\sum_{k=1}^{N}r_k\right)^2. \quad (3.2)$$

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:

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

We can similarly establish a lower bound so that

$$\frac{C_1}{N} \leq \mathbb{E}W_2^2(f_N, g_N) \leq \frac{C_2}{N}, \qquad\qquad (3.3)$$

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 to mass one in order to obtain the desired result. Recalling that $g = g/(1+\bar{r}^N)$, we can

rescale the squared Wasserstein metric [22, Proposition 7.16] to obtain

$$W_2^2(f_N/\langle f_N\rangle,g)=\left(\frac{1}{1+\bar r^N}\right)^2 W_2^2(f_N,g_N),$$

where

$$\left(\frac{1}{1+c}\right)^2\leq\left(\frac{1}{1+\bar r^N}\right)^2\leq\left(\frac{1}{1-c}\right)^2.$$

Thus we conclude that $\mathbb{E}W_2^2(f_N/\langle f_N\rangle,g)=\mathcal{O}(\frac{1}{N})$. □

REMARK 3.1.   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 (1.4), it is clear that any transport 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 distribution 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.

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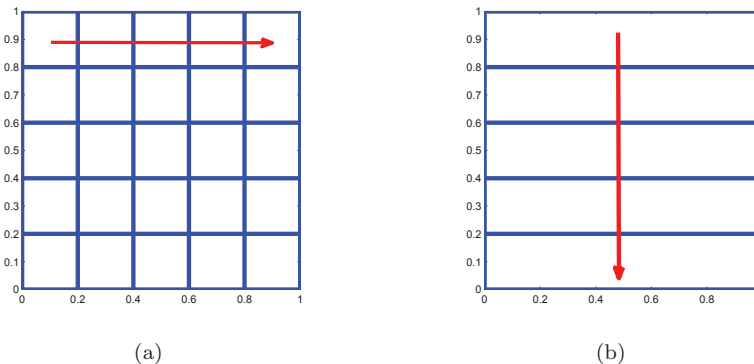

FIG. 3.3. *(a) The optimal map for each row: $T_x=T_i$ for $x_i<x\leq x_{i+1}$ and (b) the optimal map in y direction: T_y*

One approach to rearranging all the mass from f_N to g is to define T in two steps as in Figure 3.3. 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 $x_{i-1} < x \le x_i$, $i = 1, \dots, N$.

The rearrangement determined by T_N is not optimal, but does provide an upper bound on the value of the Wasserstein metric:

$$\mathbb{E}W_2^2(f_N, g) \le \mathbb{E}\left(\iint \left(|x - T_x(x)|^2 + |y - T_y(y)|^2 \right) f_N(\mathbf{x}) dx dy \right) \propto \mathcal{O}\left(\frac{1}{N} \right). \qquad (3.4)$$

Finally, by the Lebesgue dominated convergence theorem,

$$\mathbb{E}W_2^2(f, g) \le \lim_{N \to \infty} \mathbb{E}\left(\int_{\mathbb{R}^2} |\mathbf{x} - T_N(\mathbf{x})|^2 f_N(\mathbf{x}) d\mathbf{x} \right) = 0. \qquad (3.5)$$

For higher dimensions $n \ge 3$, we can similarly reduce the problem to several 1D optimal transport problems. 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 with non-homogeneous Neumann boundary conditions:

$$\begin{cases} \det(D^2 u(x)) = f(x)/g(\nabla u(x)) + \langle u \rangle, & x \in X \\ \nabla u(x) \cdot \nu = x \cdot \nu, & x \in \partial X \\ u \text{ is convex.} \end{cases} \qquad (4.1)$$

REMARK 4.1.   The Neumann boundary condition is easily generalised to the situation where f and g are supported on different rectangles [8].

The squared Wasserstein metric is then given by

$$W_2^2(f, g) = \int_X f(x) |x - \nabla u(x)|^2 \, dx. \qquad (4.2)$$

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 [8].

$$\begin{aligned} &\det^+(D^2 u) \\ &= \min_{\{v_1, v_2\} \in V} \{ \max\{u_{v_1, v_1}, 0\} \max\{u_{v_2, v_2}, 0\} + \min\{u_{v_1, v_1}, 0\} + \min\{u_{v_2, v_2}, 0\} \} \end{aligned} \qquad (4.3)$$

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

Equation (4.3) can be discretized 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 [8, 9] for complete details.

We begin by introducing the finite difference operators

$$[\mathcal{D}_{x_1x_1}u]_{ij} = \frac{1}{dx^2}\left(u_{i+1,j}+u_{i-1,j}-2u_{i,j}\right),$$

$$[\mathcal{D}_{x_2x_2}u]_{ij} = \frac{1}{dx^2}\left(u_{i,j+1}+u_{i,j-1}-2u_{i,j}\right),$$

$$[\mathcal{D}_{x_1}u]_{ij} = \frac{1}{2dx}\left(u_{i+1,j}-u_{i-1,j}\right),$$

$$[\mathcal{D}_{x_2}u]_{ij} = \frac{1}{2dx}\left(u_{i,j+1}-u_{i,j-1}\right),$$

$$[\mathcal{D}_{vv}u]_{ij} = \frac{1}{2dx^2}\left(u_{i+1,j+1}+u_{i-1,j-1}-2u_{i,j}\right),$$

$$[\mathcal{D}_{v^\perp v^\perp}u]_{ij} = \frac{1}{2dx^2}\left(u_{i+1,j-1}+u_{i+1,j-1}-2u_{i,j}\right),$$

$$[\mathcal{D}_{v}u]_{ij} = \frac{1}{2\sqrt{2}dx}\left(u_{i+1,j+1}-u_{i-1,j-1}\right),$$

$$[\mathcal{D}_{v^\perp}u]_{ij} = \frac{1}{2\sqrt{2}dx}\left(u_{i+1,j-1}-u_{i-1,j+1}\right).$$

In the compact version of the scheme, the minimum in Equation (4.3) is approximated using only two possible values. The first uses directions aligning with the grid axes.

$$MA_1[u] = \max\{\mathcal{D}_{x_1x_1}u,\delta\}\max\{\mathcal{D}_{x_2x_2}u,\delta\}$$
$$+ \min\{\mathcal{D}_{x_1x_1}u,\delta\} + \min\{\mathcal{D}_{x_2x_2}u,\delta\} - f/g(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u) - u_0. \quad (4.4)$$

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

$$MA_2[u] = \max\{\mathcal{D}_{vv}u,\delta\}\max\{\mathcal{D}_{v^\perp v^\perp}u,\delta\} + \min\{\mathcal{D}_{vv}u,\delta\} + \min\{\mathcal{D}_{v^\perp v^\perp}u,\delta\}$$
$$- f/g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_{v}u+\mathcal{D}_{v^\perp}u), \frac{1}{\sqrt{2}}(\mathcal{D}_{v}u-\mathcal{D}_{v^\perp}u)\right) - u_0. \quad (4.5)$$

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

$$M_M[u] \equiv -\min\{MA_1[u], MA_2[u]\} = 0. \quad (4.6)$$

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

$$M_N[u] \equiv -\left((\mathcal{D}_{x_1x_1}u)(\mathcal{D}_{x_2x_2}u) - (\mathcal{D}_{x_1x_2}u^2)\right) + f/g(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u) + u_0 = 0. \quad (4.7)$$

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

$$M_F[u] \equiv M_M[u] + \epsilon S\left(\frac{M_N[u] - M_M[u]}{\epsilon}\right) = 0, \quad (4.8)$$

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

$$S(x) = \begin{cases} x, & |x| \le 1, \\ 0, & |x| \ge 2, \\ -x+2, & 1 \le x \le 2, \\ -x-2, & -2 \le x \le -1. \end{cases} \tag{4.9}$$

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

$$W_2^2(f,g) \approx \sum_{j=1}^m (x_j - D_{x_j}u_h)^T \operatorname{diag}(f)(x_j - D_{x_j}u_h). \tag{4.10}$$

The computation of the discrete solution of Equation (4.8) 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

$$\nabla M_F[u] = \left(1 - S'\left(\frac{M_N[u] - M_M[u]}{\epsilon}\right)\right)\nabla M_M[u] + S'\left(\frac{M_N[u] - M_M[u]}{\epsilon}\right)\nabla M_N[u]. \tag{4.11}$$

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

$$\nabla_u M_1[u] = \left(\max\{\mathcal{D}_{x_2 x_2}, \delta\}\mathbb{1}_{\mathcal{D}_{x_1 x_1} > \delta} + \mathbb{1}_{\mathcal{D}_{x_1 x_1} \le \delta}\right)\mathcal{D}_{x_1 x_1}$$
$$+ \left(\max\{\mathcal{D}_{x_1 x_1}, \delta\}\mathbb{1}_{\mathcal{D}_{x_2 x_2} > \delta} + \mathbb{1}_{\mathcal{D}_{x_2 x_2} \le \delta}\right)\mathcal{D}_{x_2 x_2}$$
$$- \frac{f}{g(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u)^2}\nabla g(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u)\cdot(\mathcal{D}_{x_1}, \mathcal{D}_{x_2}) - \mathbb{1}_{x=x_0},$$

$$\nabla_u M_2[u] = \left(\max\{\mathcal{D}_{v^\perp v^\perp}, \delta\}\mathbb{1}_{\mathcal{D}_{vv} > \delta} + \mathbb{1}_{\mathcal{D}_{vv} \le \delta}\right)\mathcal{D}_{vv}$$
$$+ \left(\max\{\mathcal{D}_{vv}, \delta\}\mathbb{1}_{\mathcal{D}_{v^\perp v^\perp} > \delta} + \mathbb{1}_{\mathcal{D}_{v^\perp v^\perp} \le \delta}\right)\mathcal{D}_{v^\perp v^\perp}$$
$$- f/g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v u + \mathcal{D}_{v^\perp}u), \frac{1}{\sqrt{2}}(\mathcal{D}_v u - \mathcal{D}_{v^\perp}u)\right)^2$$
$$\nabla g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v u + \mathcal{D}_{v^\perp}u), \frac{1}{\sqrt{2}}(\mathcal{D}_v u - \mathcal{D}_{v^\perp}u)\right)$$
$$\cdot\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v + \mathcal{D}_{v^\perp}), \frac{1}{\sqrt{2}}(\mathcal{D}_v - \mathcal{D}_{v^\perp})\right) - \mathbb{1}_{x=x_0},$$

$$\nabla_u M_M[u] = -\mathbb{1}_{M_M[u]=-M_1[u]}\nabla_u M_1[u] - \mathbb{1}_{M_M[u]=-M_2[u]}\nabla_u M_2 u],$$

$$\nabla_u M_N[u] = -(\mathcal{D}_{x_2 x_2}u)\mathcal{D}_{x_1 x_1} - (\mathcal{D}_{x_1 x_1}u)\mathcal{D}_{x_2 x_2} + 2(\mathcal{D}_{x_1 x_2}u)\mathcal{D}_{x_1 x_2}$$
$$+ \frac{f}{g(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u)^2}\nabla g(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u)\cdot(\mathcal{D}_{x_1}, \mathcal{D}_{x_2}) + \mathbb{1}_{x=x_0}.$$

The availability of these Jacobians will become key in Section 5, where we will use these results to compute the Fréchet gradient of the Wasserstein metric. This, in turn, is needed for the minimization in the computational examples of Section 6.

5. Computation of Fréchet 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

$$\nabla_f W_2^2(f(v))\nabla_v f(v).$$

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 minimization 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 $\nabla_v f$.

Two different approaches are possible here. One option is to directly linearize the Wasserstein metric, then discretize the result. A second approach, which we pursue here, is to linearize the discrete approximation of the Wasserstein metric. A key advantage of this approach is that it allows us to make use of the Jacobian (4.11) that is already being constructed in the process of solving the Monge–Ampère equation. We also argue that this is the correct gradient since our approach to full waveform inversion is exactly solving the optimization problem (1.3) where the misfit function $d(f,g)$ is given by a discrete approximation to the squared Wasserstein metric.

Using the finite difference matrices introduced in Section 4, we can express the discrete Wasserstein metric as

$$d(f,g)=\sum_{j=1}^{n}(x_j - D_{x_j}u_f)^T \operatorname{diag}(f)(x_j - D_{x_j}u_f) \tag{5.1}$$

where the potential u_f satisfies the discrete Monge–Ampère equation

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

LEMMA 5.1 (Fréchet gradient of discrete Wasserstein metric). *The Fréchet gradient of the discretized Wasserstein metric* (5.1) *is given by*

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

Proof. The first variation of the squared Wasserstein metric as

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

Linearizing 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 (4.11). Then the gradient of the discrete squared Wasserstein metric can be expressed as

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

$\square$

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. Minimization is performed using the MATLAB function `fmincon`, equipped with the gradient described in Section 5.

The wave equation (1.1) is solved by using finite difference scheme for a defined initial wave field.

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

with the initial conditions

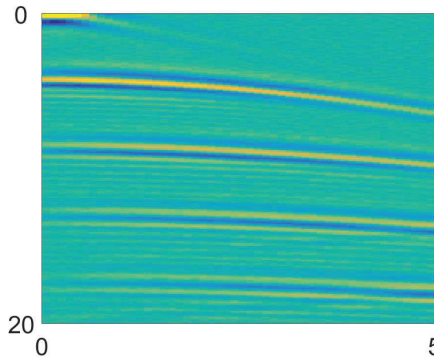$$u_{n,m}^{-1} = f(n\Delta x, m\Delta z), \quad u_{n,m}^0 = f(n\Delta x, m\Delta z).$$

FIG. 6.1. *A signal produced from a single layer model with added noise.*

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:

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

(a)   (b)
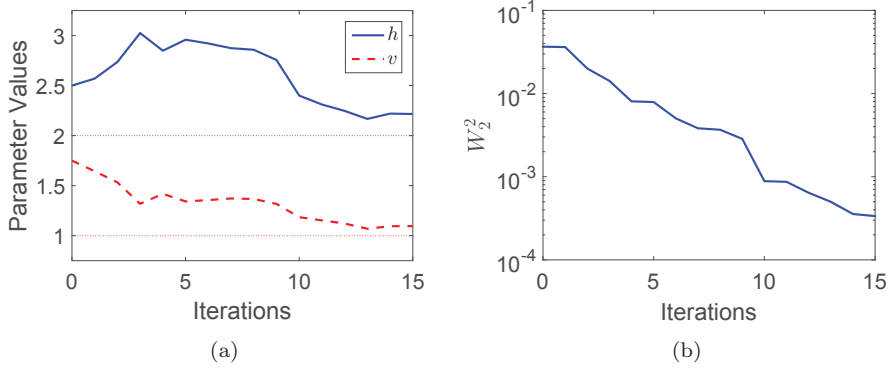
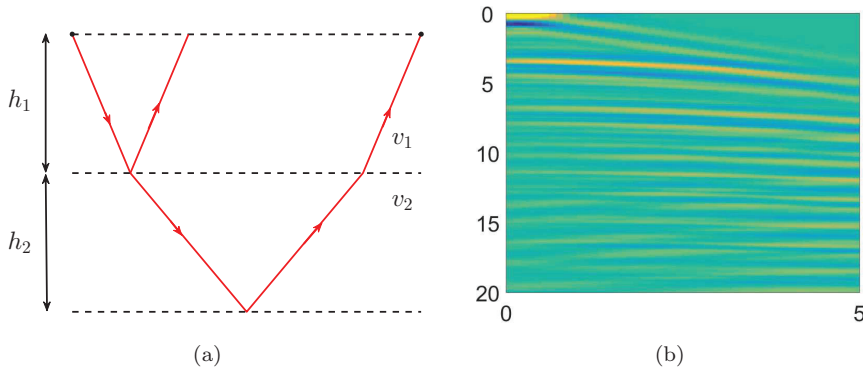FIG. 6.2. *Convergence history for a single layer model.*

(a)   (b)

FIG. 6.3. *(a) A two-layer material and (b) the resulting signal.*

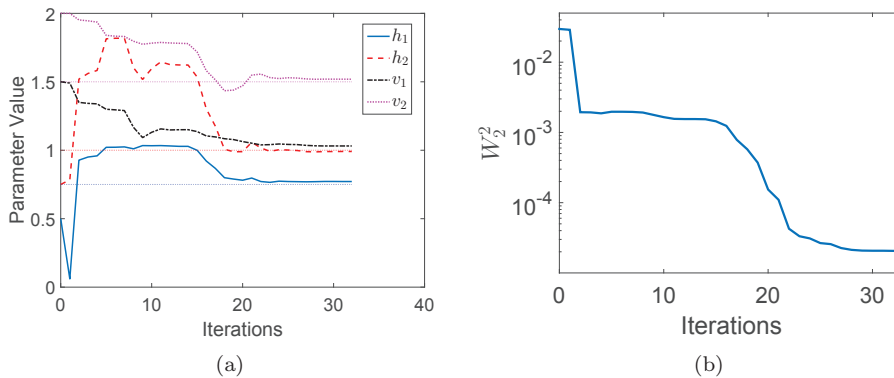(a)   (b)

FIG. 6.4. *Convergence history for a two-layer material.*

To ensure the data to be positive which is a requirement for objects in optimal transportation, we work with something akin to a local amplitude by defining

$$\tilde{f}(x,t) = \sqrt{\int_{t-\epsilon}^{t+\epsilon} u(x,0,s)^2 \, ds}$$
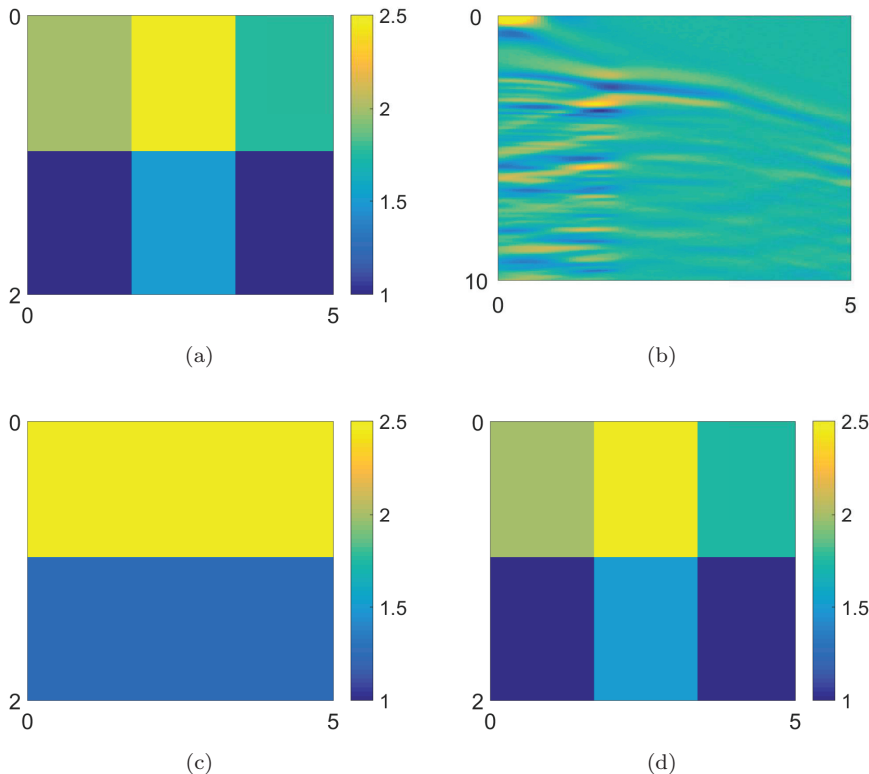
FIG. 6.5. *(a) A six parameter velocity model used to generate (b) a target signal g. (c) Initial and (d) computed velocity.*

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

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

We consider the particular case of $h^* = 2$, $v^* = 1$. 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 6.1. Then our goal is to determine h and v that minimize

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

We initialize with the guess $h = 2.5$ and $v = 1.75$ and perform minimization over the parameters h and v^{-1}. The convergence history is displayed in Figure 6.4. Despite the noise in the target profile, we recover the parameters $\tilde{h} = 2.2157$ and $\tilde{v} = 1.0953$ after fifteen iterations, with a squared Wasserstein metric of 3.36×10^{-4}. (The required step-size in the minimization algorithm became too small to improve appreciably beyond this).
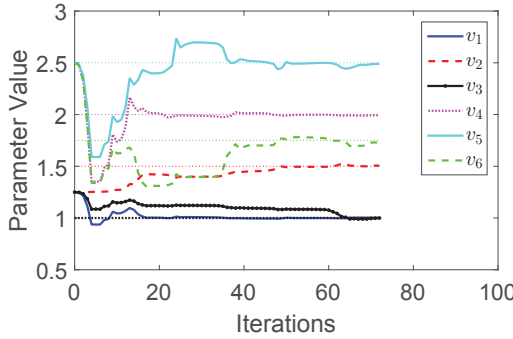
(a)
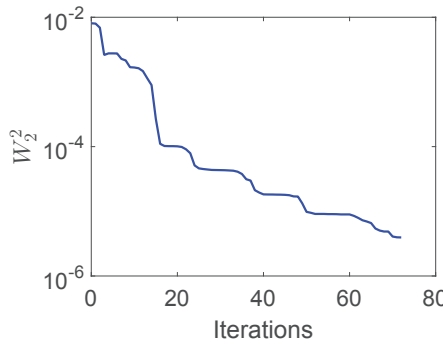
(b)

FIG. 6.6. *Convergence history for a six parameter model.*

For reference, we also compare the noisy target g with the exact signal $f_{2,1}$ (without noise). This yields a a squared Wasserstein metric of 7.49×10^{-4}, so that the error in the recovered parameters can be explained by the noise.

6.2. Two layer model.   Next, we consider the case where the material is composed of two different layers. The top layer has depth h_1 and velocity v_1 while the bottom layer has depth h_2 and velocity v_2; see Figure 6.3(a).

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

$$h_1^* = 0.75, \quad v_1^* = 1, \quad h_2^* = 1, \quad v_2^* = 1.5.$$

As in the previous example, we add noise to this target. The resulting signal is shown in Figure 6.3(b).

In this case, the distance W_2 depends on the four parameters $h_1, v_1^{-1}, h_2, v_2^{-1}$. We initialize with the guess $h_1 = 0.5$, $v_1 = 1.5$, $h_2 = 0.75$, and $v_2 = 2$. After 33 iterations, we recover the parameter values $\tilde{h}_1 = 0.772$, $\tilde{h}_2 = 0.991$, $\tilde{v}_1 = 1.0318$, and $\tilde{v}_2 = 1.519$ with a squared misfit value of 2.06×10^{-5}. The convergence history is presented in Figure 6.4.

As noted in [7], when the model involves both depth and velocity, the resulting distance can contain narrow valleys, and computing the minimum can require small step-sizes. We were still able to effectively compute the minimum in this setting, but we expect that quasi-Newton methods would enable even faster convergence.
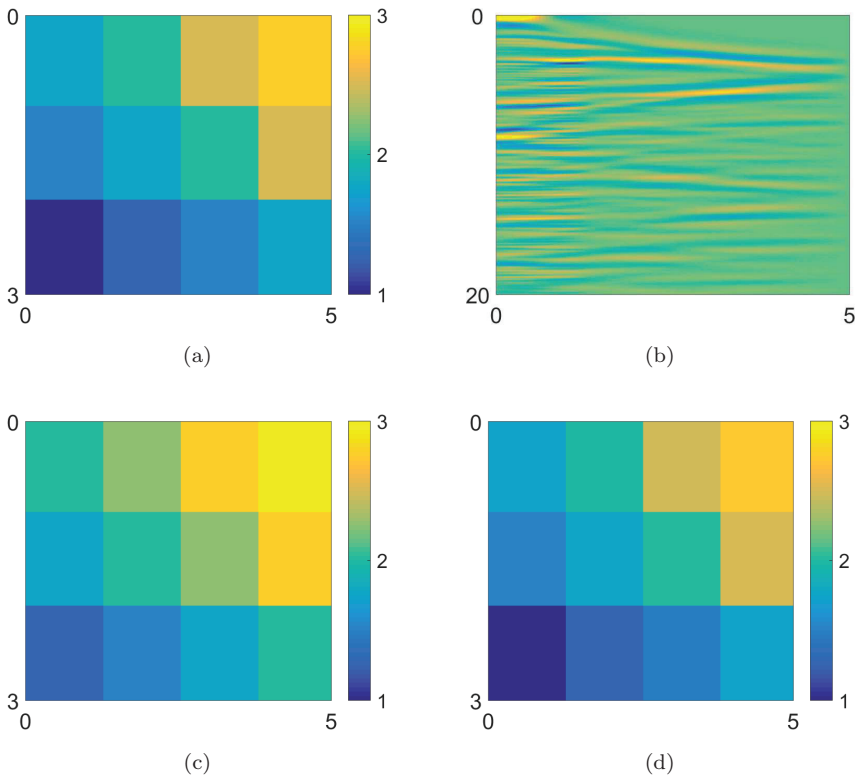
FIG. 6.7. *(a) A twelve parameter velocity model used to generate (b) a target signal g. (c) Initial and (d) computed velocity.*
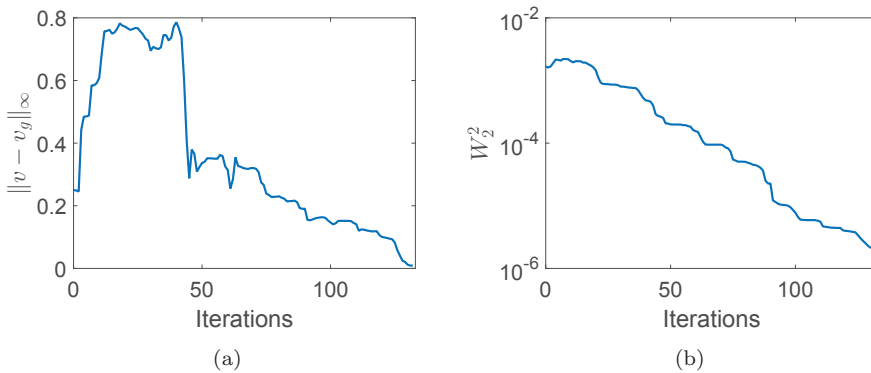

FIG. 6.8. *Convergence history for a twelve parameter model.*

6.3. Six Parameter model.   We next consider the case of a piecewise constant material. See Figure 6.5 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 6.5(b).

In this case, the distance W_2 depends on the six parameters $1/v_i$, $i = 1,\ldots,6$. We initialize 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 step-sizes that we observed in the previous example. After 72 iterations, we recover the parameter values $\tilde{v}_1 = 1.0034$, $\tilde{v}_2 = 1.5058$, $\tilde{v}_3 = 0.9996$, $\tilde{v}_4 = 1.9932$, $\tilde{v}_5 = 2.4889$, and $\tilde{v}_6 = 1.7296$ with a squared misfit value of 3.94×10^{-6}. The convergence history is presented in Figure 6.6. For reference, comparison of the noisy target with the exact signal (without noise) yielded a squared Wasserstein metric of 4.82×10^{-6}.

6.4. Twelve Parameter model. We again consider a piecewise constant velocity model, but this time increase the number of parameters to twelve. See Figure 6.7 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 initialize with the guess $v = v^* + 0.25$. After 132 iterations, we recover the twelve parameters with a maximum error of $\|\tilde{v} - v^*\|_\infty = 0.0091$ 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 6.8. The simple models in Section 6 were included to indicate how the results depend on model complexity.

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 Fréchet 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 in order to apply the method to more realistic seismic applications.

REFERENCES

[1] L. Ambrosio and N. Gigli, *A user guide to optimal transport*, in Modelling and Optimisation of Flows on Networks, Springer, 1–155, 2013.
[2] 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.
[3] J.-D. Benamou, B.D. Froese, and A.M. Oberman, *Numerical solution of the optimal transportation problem using the Monge–Ampére equation*, J. Comput. Phys., 260:107–126, 2014.
[4] D.P. Bertsekas, *Auction algorithms for network flow problems: a tutorial introduction*, Comput. Optim. Appl., 1(1):7–66, 1992.
[5] D. Bosc, *Numerical Approximation of Optimal Transport Maps*, SSRN, 2010.
[6] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44:375–417, 1991.
[7] B. Engquist and B.D. Froese, *Application of the Wasserstein metric to seismic signals*, Commun. Math. Sci., 12(5):979–988, 2014.
[8] 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.

[9] 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.

[10] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer Science & Business Media, 127, 2006.

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

[12] P. Lailly, *The seismic inverse problem as a sequence of before stack migrations*, in Conference on Inverse Scattering: Theory and Application, Society for Industrial and Applied Mathematics, Philadelphia, PA, 206–220, 1983.

[13] 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.

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

[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*, Beethoven Journal, 17(1):115-131, 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, 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 Science Publishers B.v.amsterdam, 1987.

[22] C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 58, 2003.

[23] C. Villani, *Optimal Transport: Old and New*, Springer Science & Business Media, 338, 2008.

[24] W. Zhang and J. Luo, *Full-waveform velocity inversion based on the acoustic wave equation*, American Journal of Computational Mathematics, 03(3):13-20, 2013.