Seismic Imaging and Optimal Transport

by Bjorn Engquist* and Yunan Yang†

Abstract. Seismology has changed character since 50 years ago when the full wavefield could be determined. Partial differential equations (PDE) started to be used in the inverse process of finding properties of the interior of the earth. In this paper, we will review earlier techniques focusing on Full Waveform Inversion (FWI), which is a large-scale non-convex PDE constrained optimization problem. The minimization of the objective function is usually coupled with the adjoint state method, which also includes the solution to an adjoint wave equation. The least-squares (L^2) norm is the conventional objective function measuring the difference between simulated and measured data, but it often results in the minimization trapped in local minima. One way to mitigate this is by selecting another misfit function with better convexity properties. Here we propose using the quadratic Wasserstein metric (W_2) as a new misfit function in FWI. The optimal map defining W_2 can be computed by solving a Monge-Ampère equation. Theorems pointing to the advantages of using optimal transport over L^2 norm will be discussed, and several large-scale computational examples will be presented.

AMS 2000 subject classifications: 65K10, 65K10, 86A15, 86A22.

Keywords and phrases: Seismic Imaging, Full-waveform Inversion, Optimal Transport, Monge-Ampère equation.

* Department of Mathematics and ICES, The University of Texas at Austin, 1 University Station C1200, Austin, TX 78712 USA
E-mail: engquist@math.utexas.edu
† Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York NY 10012 USA
E-mail: yunan.yang@nyu.edu

1. Introduction

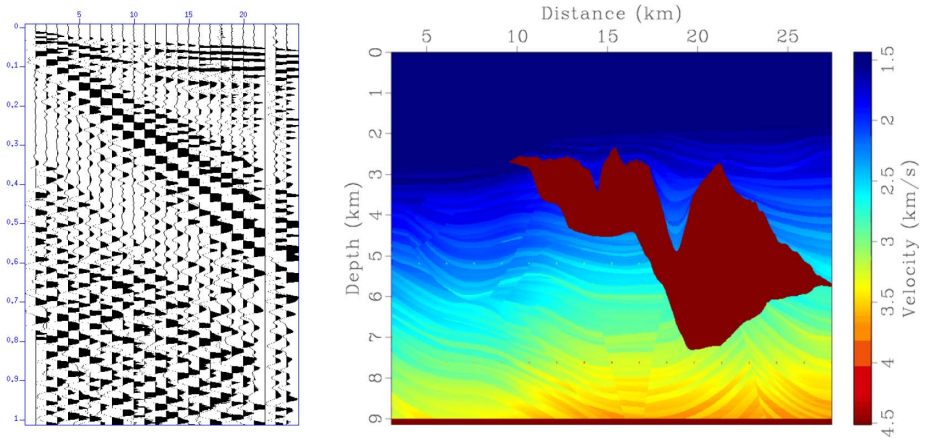
Earth Science is an early scientific subject. The efforts started as early as AD 132 in China when Heng Zhang invented the first seismoscope in the world (Figure 1a). The goal was to record that an earthquake had happened and to try to determine the direction of the earthquake. Substantial progress in seismology had to wait until about 150 years ago when seismological instruments started to record travel time.

With increasing sophistication in devices measuring the vibrations of seismic waves and in the availability of high-performance computing increasingly advances mathematical techniques could be used to explore the interior of the earth. The development started with calculations by hand based on geometrical optics and travel time measurement. It continued with a variety of wave equations when the equipment allowed for measuring wave fields and modern computers became available. As we will see below a wide range of mathematical tools are used today in seismic imaging, including partial differential equation (PDE) constrained optimization, advanced signal processing, optimal transport and the Monge-Ampère equation.

Since 19th-century modern seismographs were developed to record seismic signals, which are vibrations in the earth. In 1798 Henry Cavendish measured the density of the earth with less than 1% error compared with the number we can measure nowadays. Nearly one hundred years later, German physicist Emil Wiechert first discovered that the earth has a layered structure and the theory was further completed as the crust-mantle-core three-layer model in 1914 by one of his student Beno Gutenberg. In the meantime, people studied the waves including body

(a) (b)

Figure 1. (a) The first Seismoscope designed in AD 132 and (b) Modern seismic vibrator used in seismic survey.



(a) (b)

Figure 2. (a) An example of the seismic data measured from the receivers and (b) Goal of inversion: geophysical properties as in the Sigsbee velocity model [5].

waves and surface waves to better understand the earthquake. P-waves and S-waves were first clearly identified for their separate arrivals by English geologist Richard Dixon Oldham in 1897. The Murchison earthquake in 1929 inspired the Danish female seismologist and geophysicist Inge Lehmann to study the unexpected P-waves recorded by the seismographs. Later on, she proposed that the core of the earth has two parts: the solid inner core of iron and a liquid outer core of nickel-iron alloy, which was soon acknowledged by peer geophysicists worldwide.

We will see that measuring travel time plays a vital role in the development of modern techniques for the inverse problem of finding geophysical properties from measurements of seismic waves on the surface. The methods are often related to travel time tomography. They are quite robust and cost-efficient for achieving low-resolution information of the subsurface velocities. The forward problem is based on ray theory or geometric optics [12, 128].

The development of man-made seismic sources and advanced recording devices (Figure 1b) facilitate the research on the entire wavefields in time and space (Figure 2a) rather than merely travel time. This setup results in a more controlled setting and large amounts of data, which is needed for an accurate inverse process of estimating geophysical properties, for example, Figure 2b. The forward modeling is a wave equation with many man-made sources and many receivers. The wave equation can vary from pure acoustic waves to anisotropic viscoelasticity. Even if there are various techniques in computational exploration seismology, there are two processes that currently stand out: reverse time migration (RTM) [6, 144] and full waveform inversion (FWI) [121, 124].

Migration techniques can be applied in both the time domain and the frequency domain following the early breakthroughs by Claerbout on imaging conditions [33, 34]. In reverse time migration (RTM), the computed forward wavefield starting from the source is correlated in time with the computed backward

wavefield which is modeled with the measured data as the source term in the adjoint wave equation. The goal is to determine details of the reflecting surfaces as, for example, faults and sedimentary layers based on the measured data and a rough estimate of the geophysical properties. The least-squares reverse time migration (LSRTM) [43] is a new migration technique designed to improve the image quality generated by RTM. Reflectivity is regarded as a small perturbation in velocity, and the quantity is recovered through a linear inverse problem.

FWI is a high-resolution seismic imaging technique which recently gets great attention from both academia and industry [132]. The goal of FWI is to find both the small-scale and large-scale components which describe the geophysical properties using the entire content of seismic traces. A trace is the time history measured at a receiver. In this paper, we will consider the inverse problem of finding the wave velocity of an acoustic wave equation in the interior of a domain from knowing the Cauchy boundary data together with natural boundary conditions [36], which is implemented by minimizing the difference between computed and measured data on the boundary. It is thus a PDE-constrained optimization.

There are various kinds of numerical techniques that are used in seismic inversion, but FWI is increasing in popularity even if it is still facing three main computational challenges. First, the physics of seismic waves are complex, and we need more accurate forward modeling in inversion going from pure acoustic waves to anisotropic viscoelasticity [133]. Second, even as PDE-constrained optimization, the problem is highly non-convex. FWI requires more efficient and robust optimization methods to tackle the intrinsic nonlinearity. Third, the least-squares norm, classically used in FWI, suffers from local minima trapping, the so-called cycle skipping issues, and sensitivity to noise [114]. We will see that optimal transport based Wasserstein metric is capable of dealing with the last two limitations by including both amplitudes mismatches and travel time differences [47, 48].

We will introduce the mathematical formulation of these techniques in the following sections. The emphasis will be on FWI, but we will also summarize the state of the art of other standard imaging steps. Finally, we will relate FWI to RTM and LSRTM. These approaches all involve the interaction of the forward and the time-reversed wavefields, which is well known as the "imaging condition" in geophysics.

2. Seismic Imaging

Seismic data contains interpretable information about subsurface properties. Imaging predicts the spatial locations as well as specifies parameter values describing the earth properties that are useful in seismology. It is particularly important for exploration seismology which mainly focuses on prospecting for energy sources, such as oil, gas, coal. Seismic attributes contain both travel time records and waveform information to create an image of the subsurface to enable geological interpretation, and to obtain an estimate of the distribution of material properties in the underground. Usually, the problem is formulated as an inverse problem incorporating both physics and mathematics. Seismic inversion and migration are terms often used in this setting.

2.1 Seismic Data

There are two types of seismic signals. Natural earthquakes propagate with substantial ultra-low frequency wave energy and penetrate deeply through the whole earth. Recorded by seismometers, the natural seismic waves are used to study earth structures. The other type of data is generated by man-made "earthquakes" to obtain an image of the sedimentary basins in the interior of the earth close to the surface. A wavefield has to be produced using suitable sources at appropriate locations, measured by receivers at other locations after getting reflected back from within the earth, and stored using recorders.

In this paper, we mainly discuss the second type of seismic events. The raw seismic data is not ideal to interpret and to create an accurate image of the subsurface. Recorded artifacts are related to the surface upon which the survey was performed, the instruments of receiving and recording and the noise generated by the procedure. We must remove or at least minimize these artifacts. Seismic data processing aims to eliminate or reduce these effects and to leave only the influences due to the structure of geology for interpretation. Typical data processing steps include but are not limited to deconvolution, demultiple, deghosting, frequency filtering, normal moveout (NMO) correction, dip moveout (DMO) correction, common midpoint (CMP) stack, vertical seismic profiling (VSP), etc [108, 142].

In the recent two decades, the availability of the increased computer power makes it possible to process each trace of the recorded common source gathers separately, aiming for a better image. We will discuss several primary imaging methods such as traveltime tomography, seismic migration, least squares migration and full waveform inversion (FWI).

2.2 Traveltime Tomography

Most discoveries related to the structure of the earth were based on the assumption that seismic

waves can be represented by rays, which is closely associated with geometric optics [107, 106, 145]. The primary advantages are its applicability to complex, isotropic and anisotropic, laterally varying layered media and its numerical efficiency in such computations. A critical observation is the travel time information of seismic arrivals. We can understand many arrival time observations with ray theory [26], which describes how short-wavelength seismic energy propagates.

As a background illustration, we will derive the ray tracing expressions in a 1D setting where the velocity only varies vertically [116]. Ray tracing in general 3D structure is more complicated but follows similar principles. Considering a laterally homogeneous earth model where velocity v only depends on depth, the ray parameter which is also called the horizontal slowness p, can be expressed in the following equation by the Snell's law:

$$p = s(z)\sin(\theta) = \frac{dT}{dX}, \tag{1}$$

where $s(z)$ $(= \frac{1}{v(z)})$ is the slowness, θ is the incidence angle, T is the travel time, X is the horizontal range. At the turning point depth z_p, $p = s(z_p)$, a constant for a given ray. The vertical slowness $\eta = \sqrt{s^2 - p^2}$.

When the velocity is a continuous function of depth, the surface to surface travel time $T(p)$ and the distance traveled $X(p)$ have the following expressions:

$$T(p) = 2\int_0^{z_p} \frac{s^2(z)}{\sqrt{s^2(z) - p^2}} dz = 2\int_0^{z_p} \frac{s^2(z)}{\eta} dz, \tag{2}$$

and

$$X(p) = 2p\int_0^{z_p} \frac{dz}{\sqrt{s^2(z) - p^2}} = 2p\int_0^{z_p} \frac{dz}{\eta}. \tag{3}$$

The expressions above are the forward problem in traveltime tomography. The seismologists are interested in inverting model parameter $s(z)$ from observed traveltime T and traveled distance X. Using integral transform pair, we can obtain

$$z(s) = -\frac{1}{\pi}\int_{s_0}^{s} \frac{X(p)}{\sqrt{p^2 - s^2(z)}} d(p) = \frac{1}{\pi}\int_0^{X(s)} \cosh^{-1}(p/s) dX, \tag{4}$$

which gives us the 1D velocity model.

Equation (4) is one example of the 1D velocity inversion problem at a given depth. There are limitations about traveltime tomography in general. First, the first arrivals are inherently nonunique. Second, the lateral velocity variations are not considered in this setting. If we divide the earth model into blocks, the 3D velocity inversion techniques can resolve some of the lateral velocity perturbations by using the travel time in each block. The problem can be formulated into a least-squares (L^2) inversion by minimizing the travel time residual between the predicted time and the observed time: $||t_{\text{obs}} - t_{\text{pred}}||_2^2$ [116, 146].

One limitation of ray theory is that it is applicable only to smooth media with smooth interfaces, in which the characteristic dimensions of inhomogeneities are considerably larger than the dominant wavelength of the considered waves. The ray method can yield distorted results and will fail at caustics or in general at so-called singular regions [28]. Moreover, much more information is available from the observed seismograms than travel times. To some extent, travel time tomography can be seen as phase-based inversion, and next, we will introduce waveform-based methods where the wave equation plays a significant role.

2.3 Reverse Time Migration

To overcome the difficulties of ray theory and further improve image resolutions, reverse time migration (RTM), least-squares reverse time migration (LSRTM) and full-waveform inversion (FWI) replace the semi-analytical solutions to the wave equation by fully numerical solutions including the full wavefield. Without loss of generality, we will explain all the methods in a simple acoustic setting:

$$\begin{cases} m(\mathbf{x})\frac{\partial^2 u(\mathbf{x},t)}{\partial t^2} - \triangle u(\mathbf{x},t) = s(\mathbf{x},t) \\ u(\mathbf{x},0) = 0 \\ \frac{\partial u}{\partial t}(\mathbf{x},0) = 0 \end{cases} \tag{5}$$

We assume the model $m(\mathbf{x}) = \frac{1}{c(\mathbf{x})^2}$ where $c(\mathbf{x})$ is the velocity, $u(\mathbf{x},t)$ is the wavefield, $s(\mathbf{x},t)$ is the source. It is a linear PDE but a nonlinear operator from model domain $m(\mathbf{x})$ to data domain $u(\mathbf{x},t)$.

Despite the fact that migration can be used to update velocity model [80, 110, 119], its chief purpose is to transform measured reflection data into an image of reflecting interfaces in the subsurface. There are two principal varieties of migration techniques: reverse time migration (RTM) which gives a modest resolution of the reflectivity [6, 143] and least-squares reverse-time migration (LSRTM) which typically yields a higher resolution of the reflectivity [43, 44].

Reverse-time migration is a prestack two-way wave-equation migration to illustrate complex structure, especially strong contrast geological interfaces such as environments involving salts. Conventional RTM uses an imaging condition which is the zero time-lag cross-correlation between the source and the receiver wavefields [33]:

$$R(\mathbf{x}) = \sum_{\text{shots}} \int_0^T u(\mathbf{x},t) \cdot v(\mathbf{x},t) dt, \tag{6}$$

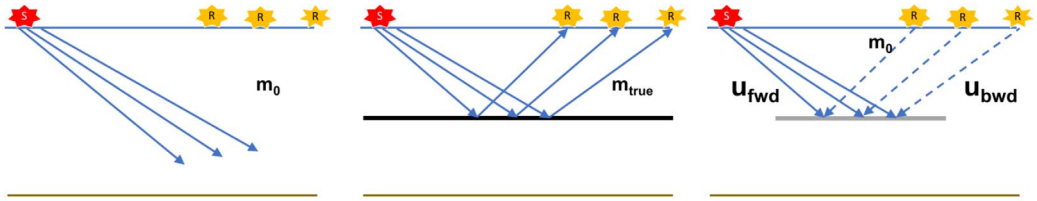


Figure 3. RTM: (a) Synthetic forward wavefield u_{fwd}, (b) True forward wavefield and (c) Reflectors generated as the backward wavefield u_{bwd} cross-correlated with u_{fwd}.

where u is the source wavefield in (5) and v is the receiver wavefield which is the solution to the adjoint equation (7):

$$(7)\qquad \begin{cases} m(\mathbf{x})\frac{\partial^2 v(\mathbf{x},t)}{\partial t^2} - \triangle v(x,t) = d(\mathbf{x},t)\delta(\mathbf{x}-\mathbf{x_r}) \\ v(\mathbf{x},T) = 0 \\ v_t(\mathbf{x},T) = 0 \end{cases}$$

Here T is the final recording time, d is the observed data from the receiver $\mathbf{x_r}$ and m is the assumed background velocity. The adjoint wave equation (7) is always solved backward in time from T to 0. Therefore it is also referred as backward propagation.

In classical RTM, the forward modeling typically does not contain reflection information. For example, it can be the paraxial approximation of the wave equation, which does not allow for reflections [36], or a smooth velocity model with unknown reflecting layers. As a summary, the conventional RTM consists three steps as Figure 3 shows:

1. Forward modeling of a wave field with a good velocity model to get u_{fwd};
2. Backpropagation of the measured data through the same model to get u_{bwd};
3. Cross-correlation the source wavefield u_{fwd} and receiver wavefield u_{bwd} based on an imaging condition (e.g., Equation (6)) to detect the reflecting interfaces.

RTM uses the entire solution of the wave equations instead of separating the downgoing or upgoing wavefields. Theoretically, RTM produces a more accurate image than ray-based methods since it does not rely on the asymptotic theory or migration using the one-way equation, which typically introduces modeling errors [113]. A good background velocity model that contains accurate information about the low-wavenumber components is also crucial for the quality of the image [55]. Recent advances in computation power make it possible to compute and store the solution of the wave equation efficiently, which significantly aids RTM to generate high-quality images [49].

2.4 Least-Squares Reverse Time Migration

Least-squares reverse time migration (LSRTM) is a new migration method designed to improve the image quality generated by RTM. It is formulated as a linear inverse problem based on the Born approximation which we will describe briefly in this section. The wave equation (5) defines a nonlinear operator $\mathcal{F}$ from model domain to data domain that maps m to u. The Born approximation is a linearization of this map to the first order so that we can denote it as $L = \frac{\delta \mathcal{F}}{\delta m}$ [64, 129].

One can derive the Born approximation as follows [46]. If we denote the model $m(\mathbf{x})$ as the sum of a background model and a small perturbation:

$$(8)\qquad m(\mathbf{x}) = m_0(\mathbf{x}) + \varepsilon m_1(\mathbf{x}),$$

the corresponding wavefield u also splits into two parts:

$$(9)\qquad u(\mathbf{x},t) = u_0(\mathbf{x},t) + u_{sc}(\mathbf{x},t),$$

where u satisfies (5), and u_0 solves the following equation:

$$(10)\qquad \begin{cases} m_0(\mathbf{x})\frac{\partial^2 u_0(\mathbf{x},t)}{\partial t^2} - \triangle u_0(\mathbf{x},t) = s(\mathbf{x},t) \\ u_0(\mathbf{x},0) = 0 \\ \frac{\partial u_0}{\partial t}(\mathbf{x},0) = 0 \end{cases}$$

Subtracting (10) from (5) and using (8), we derive an equation of u_{sc} with zero initial conditions:

$$(11)\qquad m_0\frac{\partial^2 u_{sc}(\mathbf{x},t)}{\partial t^2} - \triangle u_{sc}(\mathbf{x},t) = -\varepsilon m_1 \frac{\partial^2 u(\mathbf{x},t)}{\partial t^2}.$$

We can write u_{sc} using Green's function G:

$$(12)\qquad u_{sc}(\mathbf{x},t) = -\varepsilon \int_0^t \int_{\mathbb{R}^n} G(\mathbf{x},y;t-s) m_1(y) \frac{\partial^2 u}{\partial t^2}(y,s)\,dy\,ds.$$

As a result, the original wavefield u has an implicit relation:

$$(13)\qquad u = u_0 - \varepsilon G m_1 \frac{\partial^2 u}{\partial t^2} = \left[I + \varepsilon G m_1 \frac{\partial^2}{\partial t^2}\right]^{-1} u_0$$

The last term can be expanded in terms of Born series,

$$u = u_0 - \varepsilon \int_0^t \int_{\mathbb{R}^n} G(\mathbf{x}, y; t-s) m_1(y) \frac{\partial^2 u_0}{\partial t^2}(y,s) dy ds + \mathcal{O}(\varepsilon^2) \tag{14}$$

$$= u_0 + \varepsilon u_1 + \mathcal{O}(\varepsilon^2) \tag{15}$$

Therefore, we can approximate u_{sc} explicitly by εu_1 as $-\varepsilon G m_1 \frac{\partial^2 u_0}{\partial t^2}$, which is called the Born approximation. We also derive a linear map from m_1 to u_1:

$$\begin{cases} m_0 \frac{\partial^2 u_1(\mathbf{x},t)}{\partial t^2} - \triangle u_1(\mathbf{x},t) = -m_1 \frac{\partial^2 u_0(\mathbf{x},t)}{\partial t^2} \\ u_1(\mathbf{x},0) = 0 \\ \frac{\partial u_1}{\partial t}(\mathbf{x},0) = 0 \end{cases} \tag{16}$$

Unlike (11), (16) is an explicit formulation with m_0 as the background velocity and u_0 as the background wavefield which is the solution to (10).

It is convenient to denote the nonlinear forward map (5) as $\mathcal{F} : m \mapsto u$. A Taylor expansion of $u = \mathcal{F}(m)$ in the sense of calculus of variation, gives us:

$$u = u_0 + \varepsilon \frac{\delta \mathcal{F}}{\delta m}[m_0] m_1 + \frac{\varepsilon^2}{2} < \frac{\delta^2 \mathcal{F}}{\delta m^2}[m_0] m_1, m_1 > + \dots \tag{17}$$

The functional derivative $\frac{\delta \mathcal{F}}{\delta m} : m_1 \mapsto u_1$ is the linear operator (16), which we hereafter denote as L. The convergence of the Born series and the accuracy of the Born approximation can be proved mathematically [95, 96].

We assume there is an accurate background velocity model m_0. The Born modeling operator maps the reflectivity m_r to the scatted wavefield $d_r = \mathcal{F}(m) - \mathcal{F}(m_0)$:

$$L m_r = d_r \tag{18}$$

Although L is linear, there is no guarantee that it is invertible [35]. Instead of computing L^{-1}, we seek the reflectivity model by minimizing the least-squares error between observed data d_r and predicted scattering wavefield:

$$J(m_r) = ||L m_r - d_r||_2^2 \tag{19}$$

The normal least-squares solution to (19) is $m_r = (L^T L)^{-1} L^T d_r$ where L^T is the adjoint operator, but it is numerically expensive and unstable to invert the term $L^T L$ directly. Instead, the problem is solved in an iterative manner using optimization methods such as conjugate gradient descent (CG).

Another interesting way of approximating $(L^T L)^{-1}$ is to consider the problem as finding a non-stationary matching filter [56, 61]. Similar to RTM, we can get an image by doing one step of migration:

$$m_1 = L^T d_r. \tag{20}$$

One step of de-migration (Born modeling) based on m_1 generates data d_1

$$d_1 = L m_1. \tag{21}$$

Finally, the re-migration step provides another image m_2

$$m_2 = L^T d_1. \tag{22}$$

Combining (20) to (22), the inverse Hessian operator $(L^T L)^{-1}$ behaves like a matching filter between m_1 and m_2 which we are able to produce from the observed data. It is also the filter between m_r and m_1 as (23) and (24) show below:

$$m_1 = (L^T L)^{-1} m_2 \tag{23}$$

$$m_r = (L^T L)^{-1} m_1 \tag{24}$$

Therefore, LSRTM can be seen as a process which first derives a filter to match the re-migration m_2 to the initial migration m_1 and then applies the filter back to the initial migrated image to give an estimate of the reflectivity. Seeking the reflectivity is equivalent to finding the best filter K by minimizing the misfit $J(K)$ in the image or model domain:

$$J(K) = ||m_1 - K m_2||_2^2. \tag{25}$$

The final reflectivity image $m_r \approx K m_1$. It is a single-iteration method which greatly reduces the computational cost of the iterative methods like CG.

A potentially better way of implementing the filter-based idea is to transform the image into curvelet domain [25] to improve the stability and structural consistency in the matching [134]. The formulation of obtaining the Hessian filter in curvelet domain is to minimize a misfit function $J(s)$ where

$$J(s) = ||C(m_1) - s C(m_2)||_2^2 + \varepsilon ||s||_2^2, \tag{26}$$

where C is the curvelet domain transform operator, s is the matching filter and ε is the Tikhonov regularization parameter. The final reflectivity image $m_r \approx C^{-1}(|s| C(m_1))$, where C^{-1} is the inverse curvelet transform operator.

In general, least-squares reverse time migration (LSRTM) is still facing challenges. First of all, the image quality highly depends on the accuracy of the background velocity model m_0. Even a small error can make the two wavefields meet at a wrong location, which generates a blurred image or an incorrect reflectivity [83]. Another drawback is its high computational cost compared with other traditional migration techniques. In practice, LSRTM fits not only the data but also the noise in the data. Consequently, it boosts the high-frequency noise in the image during the iterative inversion [42, 147].

2.5 Inversion

The process of imaging through modeling the velocity structure is a form of inversion of seismic data [125], but in this paper, we regard inversion as a process of recovering the quantitative features of the geographical structure, that is, finding $m(\mathbf{x})$ in (5). Inversion is often used to build a velocity model iteratively until the synthetic data matches the actual recording [94].

Wave equation traveltime tomography [84] and the ray-based tomography in the earlier section are phase-like inversion methods [113]. Least-squares inversion is known as linearized waveform inversion [75, 122]. The migration method introduced earlier, LSRTM, can also be seen as a linear inverse problem. The background model m_0 is not updated after each iteration in least-squares inversion. Similar to the goal of migration, the model to be updated iteratively is the reflectivity distribution instead of the velocity model. One can interpret the process as a series of reverse time migrations, where the data residual is backpropagated into the model instead of the recorded data itself (Figure 3c).

If the background model m_0 is the parameter we invert for, the problem turns into a nonlinear waveform inversion, which is also called full-waveform inversion (FWI). Both the low-wavenumber and high-wavenumber components are updated simultaneously in FWI so that the final image has high resolution and high accuracy [133]. FWI is the primary focus of the paper. In the following sections, we will further discuss the topic and especially the merit of using optimal transport based ideas to tackle the current limitations.

3. Full Waveform Inversion

FWI is a nonlinear inverse technique that utilizes the entire wavefield information to estimate the earth properties. The notion of FWI was first brought up three decades ago [74, 124] and has been actively studied as the computing power increases. As we will see, the mathematical formulation of FWI is PDE constrained optimization. Even inversion for subsurface elastic parameters using FWI has become increasingly popular in exploration applications [17, 93, 133]. Currently, FWI can achieve stunning clarity and resolution. Both academia and industry have been actively working on the innovative algorithms and software of FWI. However, this technique is still facing three main challenges.

First, the physics of seismic waves are complex, and we need more accurate forward modeling in inversion going from pure acoustic waves to anisotropic viscoelasticity. Recent developments focus on this multiparameter and multi-mode modeling. FWI strategies for simultaneous and hierarchical velocity and attenuation inversion have been investigated recently [105], but there is a dilemma. The more realistic with more parameters the models of the earth become, the more ill-posed and even non-unique will the inverse problem be.

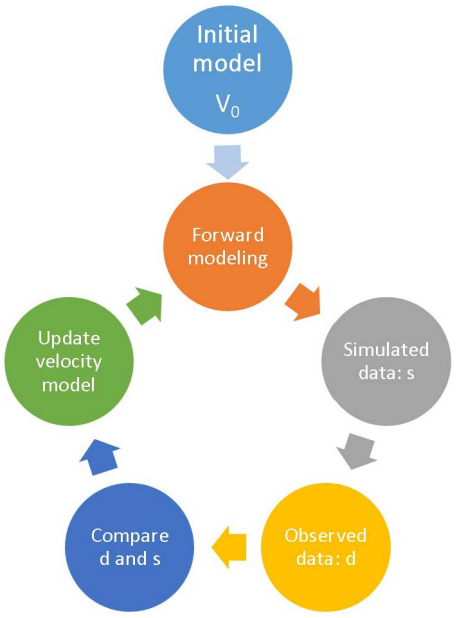


Figure 4. The framework of FWI as a PDE-constrained optimization.

Second, it is well known that the accuracy of FWI deteriorates from the lack of low frequencies, data noise, and poor starting model. The limitation is mainly due to the ill-posedness of the inverse problem which we treat as a PDE-constrained optimization. FWI is typically performed using local optimization methods in which the subsurface model is described by using a large number of unknowns, and the number of model parameters is determined a priori [123]. These methods typically only use the local gradient of the objective function. As a result, the inversion process is easily trapped in the local minima. Markov chain Monte Carlo (MCMC) based methods [109], particle swarm optimization [29], and many other global optimization methods [115] can avoid the pitfall theoretically, but they are not cost-efficient to handle practical large-scale inversion currently.

Third, it is relatively inexpensive to update the model through local optimization methods in FWI, but the convergence of the algorithm highly depends on the choice of a starting model. The research directions can be grouped into two main ideas to tackle this problem. One idea is to replace the conventional

least-squares norm with other objective functions in optimization for a wider basin of attraction [48]. The other idea is to expand the dimensionality of the unknown model by adding non-physical coefficients. The additional coefficients may convexify the problem and fit the data better [13, 62].

The essential elements of FWI framework (Figure 4) includes forward modeling and the adjoint-state method for gradient calculation.

3.1 Forward Modeling

Wave-propagation modeling is the most significant step in seismic imaging. The earth is complex with various heterogeneity on many scales, and the real physics is far more complicated than the simple acoustic setting of this paper, but the industry standard is still the acoustic model in time or frequency domain. The current research of FWI covers multiple parameters inversion of seismic waveforms including anisotropic parameters, density, and attenuation factors [138] including viscoelastic modeling which is related to fractional Laplacian wave equations [111]. It should be noted that the more parameters in a model, the less well-posed is the inverse problem.

If we exclude the attenuation parameter, the general elastic wave equation is a realistic model. Based on the equation of conservation of momentum (Newton's law of dynamics) and Hooke's law for stress and strain tensors, we have the following elastic wave equation:

$$\rho \frac{\partial^2 u_i}{\partial t^2} = f_i + \frac{\partial \sigma_{ij}}{\partial x_j}, \tag{27}$$

$$\frac{\partial \sigma_{ij}}{\partial t} = c_{ijkl} \frac{\partial \varepsilon_{ij}}{\partial t} + \frac{\partial \widetilde{\sigma}_{ij}}{\partial t}, \tag{28}$$

where ρ is the density, $\mathbf{u}$ is the displacement vector, σ is the nine-component stress tensor ($i,j = 1,2,3$), $\widetilde{\sigma}$ is the internal stress, $\mathbf{f}$ is the outer body force, ε is the nine-component strain tensor which satisfies $\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$ and c_{ijkl} is the stiffness tensor containing twenty-one independent components.

One can classify the current numerical methods of complex wave propagation into three categories: direct methods, integral wave equation methods and asymptotic methods [65]. Direct methods include finite-difference method (FDM) [91], pseudospectral method [53], finite element method (FEM) [86], spectral element method (SEM) [71], discontinuous Galerkin method (DG) [67], etc. Integral wave equation methods include both boundary element method (BEM) [14] and the indirect boundary element methods (IBEM) [101] with a fast multipole method (FMM) [52] for efficiency. Asymptotic methods include geometrical optics, Gaussian beams [27] and frozen Gaussian beams [81].

3.2 Measure of Mismatch

In seismic inversion, the misfit function, i.e. the objective function in the optimization process, is defined as a functional on the data domain. Common misfit functions include cross-correlation traveltime measurements [84, 87], amplitude variations [41] and waveform differences [124]. In both time [121] and frequency domain [102, 103], the least-squares norm has been the most widely used misfit function. For example, in time domain conventional FWI defines a least-squares waveform misfit as

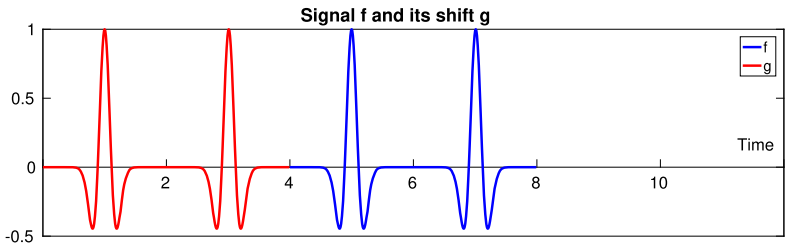
$$d(f,g) = J(m) = \frac{1}{2}\sum_r \int |f(\mathbf{x_r},t;m) - g(\mathbf{x_r},t)|^2 \, dt, \tag{29}$$

where $\mathbf{x_r}$ are receiver locations, g is observed data, and f is simulated data which solves (5) with model parameter m. The time integral is carried out numerically as a sum. This formulation can also be extended to the case with multiple sources.

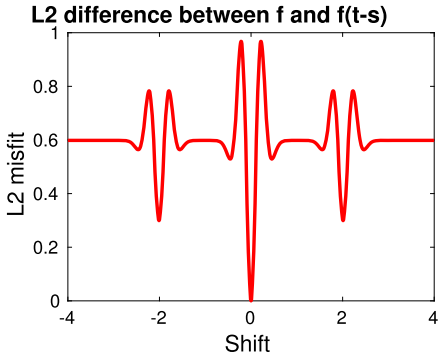
Real seismic data usually contains noise. As a result, denoising becomes an important step in seismic data processing. The L^2 norm is well known to be sensitive to noise [18]. Other norms have been proposed to mitigate this problem. For example, the L^1 norm [37, 121], the Huber criterion [57, 59] and the hybrid L^1/L^2 criterion [19] all demonstrated improved robustness to noise compared with conventional L^2 norm.

All the misfit functions above are point-by-point based objective functions which means they only accumulate the differences in amplitude at each fixed time grid point. There are global misfit functions that compare the simulated and measured signals not just pointwise. The Wasserstein metric is one such metric which we will discuss later. It is very robust with respect to noise

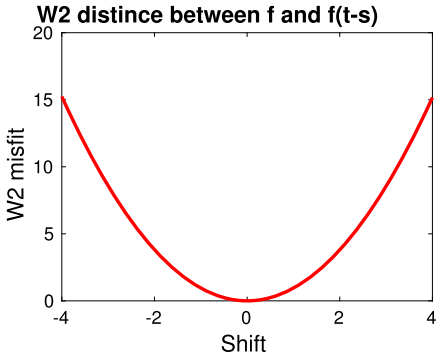
The oscillatory and periodic nature of waveforms lead to another main challenge in FWI: the cycle-skipping issue when implementing FWI as a local inversion scheme. If the true data and the initial synthetic data are more than half wavelength ($> \frac{\lambda}{2}$) away from each other, the first gradient can go in the wrong direction regarding the phase mismatch, but can nonetheless reduce the data misfit in the fastest manner [11]. Mathematically, it is related to the highly nonconvex and highly nonlinear nature of the inverse problem and results in finding only a local minima. Figure 5a displays two signals, each of which contains two Ricker wavelets and f is simply a shift of g. The L^2 norm between f and g is plotted in Figure 5b as a function of the shift s. We observe many local minima and maxima in this simple two-event setting which again demonstrated the difficulty of the, so called, cycle-skipping issues [139].



(a) Two signals

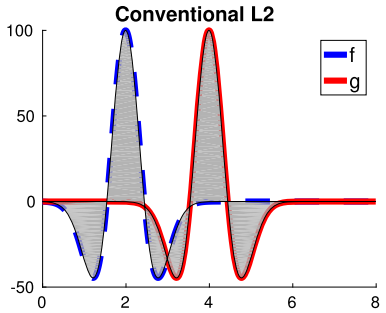


(b) L^2 sensitivity curve

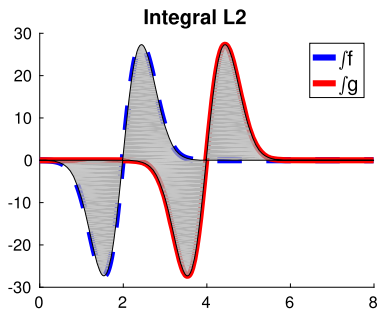


(c) W_2 sensitivity curve

Figure 5. (a) A signal consisting two Ricker wavelets (blue) and its shift (red) (b) L^2 norm of the difference between f and $f(t-s)$ in terms of shift s (c) W_2 norm between f and $f(t-s)$ in terms of shift s.



(a) Misfit for L^2 norm



(b) Misfit for Integral L^2 method

Figure 6. The shaded areas represent the mismatch each misfit function considers. (a) L^2: $\int (f-g)^2 dt$. (b) Integral wavefields method: $\int (\int f - \int g)^2 dt$. [140].

The lower frequency components have a wider basin of attraction with the least-squares norm being the misfit function. Several hierarchical methods that invert from low frequencies to higher frequencies have been proposed in the literature to mitigate the cycle-skipping of the inverse problem [20, 69, 103, 117, 137]. Several other methods instead compare the integrated waveforms [63, 79] (Figure 6) and the waveform envelops [15, 82]. They share a similar idea with the hierarchical methods of taking advantage of the lower frequency components in the data.

A recently introduced class of misfit functions is based on optimal transport [30, 47, 48, 88, 89, 139, 140, 141]. As a useful tool from the theory of optimal transport, the Wasserstein metric computes the minimal cost of rearranging one distribution into another. The optimal transport based methods compare the observed and simulated data globally and thus include phase information. We will discuss these measures in section 4 and 5.

Other misfit functions with the idea of non-local comparison proposed in the literature include filter based misfit functions [136, 148] as well as inversion using, so called, dynamic time warping [85] and the registration map [3]. The differential semblance optimization [120] exploits both phase and ampli-

tude information of the reflections. Tomographic full waveform inversion [13] has some global convergence characteristics of wave-equation migration velocity analysis. In the filter based methods [136, 148], a filter is designed to minimize the L^2 difference between filtered simulated data and the observed data. The misfit is then a measure of how much the filter deviates from the identity. As we will see in the optimal transport based technique, this is done in one step where the optimal map directly determines the mapping of the simulated data. The optimal transport map is general and does not need to have the form of a convolution filter as in the filter based methods.

3.3 Adjoint-State Method

Large-scale realistic 3D inversion is possible today. The advances in numerical methods and computational power allow for solving the 3D wave equations and compute the Fréchet derivative with respect to model parameters, which are needed in the optimization. In the adjoint-state method, one only needs to solve two wave equations numerically, the forward propagation and the backward adjoint wavefield propagation. Different misfit functions typically only affect the source term in the adjoint wave equation [100, 123].

Let us consider the misfit function $J(m)$ for computing the difference between predicted data f and observed data g where m is the model parameter, $F(m)$ is the forward modeling operator, $u(\mathbf{x},t)$ is the wavefield and $s(\mathbf{x},t)$ is the source. The predicted data f is the partial Cauchy boundary data of u which can be written as $f = Ru$ where R is a restriction operator only at the receiver locations. The wave equation (5) can be denoted as

$$F(m)u = s. \tag{30}$$

Taking first derivative regarding model m on both sides gives us:

$$\frac{\partial F}{\partial m}u + F\frac{\partial u}{\partial m} = 0. \tag{31}$$

Therefore,

$$\frac{\partial f}{\partial m} = -RF^{-1}\frac{\partial F}{\partial m}u. \tag{32}$$

By the chain rule, the gradient of misfit function J with respect to m is

$$\frac{\partial J}{\partial m} = \left(\frac{\partial f}{\partial m}\right)^T \frac{\partial J}{\partial f} \tag{33}$$

We can derive the following equation by plugging (32) into (33):

$$\frac{\partial J}{\partial m} = -u^T \left(\frac{\partial F}{\partial m}\right)^T F^{-T} R^T \frac{\partial J}{\partial f} \tag{34}$$

Equation (34) is the adjoint-state method. The term $F^{-T}R^T\frac{\partial J}{\partial f}$ denotes the backward wavefield v generated by the adjoint wave equation whose source is the data residual $R^T\frac{\partial J}{\partial f}$. The gradient is similar to the usual imaging condition (6):

$$\frac{\partial J}{\partial m} = -\int_0^T \frac{\partial^2 u(\mathbf{x},t)}{\partial t^2} v(\mathbf{x},t)dt, \tag{35}$$

where v is the solution to the adjoint wave equation:

$$\begin{cases} m\frac{\partial^2 v(\mathbf{x},t)}{\partial t^2} - \triangle v(\mathbf{x},t) = R^T\frac{\partial J}{\partial f} \\ v(\mathbf{x},T) = 0 \\ v_t(\mathbf{x},T) = 0 \end{cases} \tag{36}$$

Therefore F^T can be seen as the backward modeling operator which is similar to the adjoint wave equation (7) but with a different source term.

There are many other equivalent ways to formulate the adjoint-state method. One can refer to [46, 100] for more details.

In FWI, our aim is to find the model parameter m^* that minimizes the objective function, i.e. $m^* = \operatorname{argmin} J(m)$. For this PDE-constrained optimization, one can use the Fréchet derivative in a gradient-based iterative scheme to update the model m, such as steepest descent, conjugate gradient descent (CG), L-BFGS, Gauss-Newton method, etc. One can also derive the second-order adjoint equation for the Hessian matrix and use the full Newton's method in each iteration, but it is not practical regarding memory and current computing power. It is one of the current research interests to analyze and approximate the Hessian matrix in optimization [132].

4. Optimal Transport for FWI

Optimal transport has become a well-developed topic in mathematics since it was first brought up by Monge [92] in 1781. Due to its ability to incorporate both intensity and spatial information, optimal transport based metrics for modeling and signal processing have recently been adopted in a variety of applications including image retrieval, cancer detection, and machine learning [70]. In computer science, the metric is often called the "Earth Mover's Distance" (EMD).

The idea of using optimal transport for seismic inversion was first proposed in [47]. The Wasserstein metric is a concept based on optimal transportation [131]. Here, we transform our datasets of seismic signals into density functions of two probability distributions. Next, we find the optimal map between these two datasets and compute the corresponding transport cost as the misfit function in FWI. In this paper, we will focus on the quadratic cost function. The corresponding misfit is the quadratic Wasserstein metric (W_2). As Figure 5c shows, the convexity

of W_2 is much better than the L^2 norm when comparing oscillatory seismic data with respect to shift.

Following the idea that changes in velocity cause a shift or "transport" in the arrival time, [48] demonstrated the advantageous mathematical properties of the quadratic Wasserstein metric (W_2) and provided rigorous proofs that laid a solid theoretical foundation for this new misfit function. We can apply W_2 as misfit function in two different ways: trace-by-trace comparison which is related to 1D optimal transport in the time dimension, and the entire dataset comparison in multiple dimensions. We will see that solving the Monge-Ampère equation in each iteration of FWI is a useful technique [141] for calculating the Wasserstein distance. An analysis of the 1D optimal transport approach and the conventional misfit functions such as L^2 norm and integral L^2 norm illustrated the intrinsic advantages of this transport idea [140].

4.1 Wasserstein Metric

Let X and Y be two metric spaces with nonnegative Borel measures μ and ν respectively. Assume X and Y have equal total measure:

$$\int_X d\mu = \int_Y d\nu \tag{37}$$

Without loss of generality, we will hereafter assume the total measure to be one, i.e., μ and ν are probability measures.

Definition 1 (Mass-preserving map). *A transport map $T : X \to Y$ is mass-preserving if for any measurable set $B \in Y$,*

$$\mu(T^{-1}(B)) = \nu(B) \tag{38}$$

If this condition is satisfied, ν is said to be the push-forward of μ by T, and we write $\nu = T_{\#}\mu$.

In another word, given two nonnegative densities $f = d\mu$ and $g = d\nu$, we are interested in the mass-preserving map T such that $f = g \circ T$. The transport cost function $c(x,y)$ maps pairs $(x,y) \in X \times Y$ to $\mathbb{R} \cup \{+\infty\}$, which denotes the cost of transporting one unit mass from location x to y. The most common choices of $c(x,y)$ include $|x-y|$ and $|x-y|^2$, which denote the Euclidean norms for vectors x and y hereafter. Once we find a mass-preserving map T, the cost corresponding to T is

$$I(T,f,g,c) = \int_X c(x,T(x))f(x)\,dx.$$

While there are many maps T that can perform the relocation, we are interested in finding the optimal map that minimizes the total cost

$$I(f,g,c) = \inf_{T\in\mathcal{M}} \int_X c(x,T(x))f(x)\,dx,$$

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

Thus we have informally defined the optimal transport problem, the optimal map as well as the optimal cost, which is also called the Wasserstein distance:

Definition 2 (The Wasserstein distance). *We denote by $\mathscr{P}_p(X)$ the set of probability measures with finite moments of order p. For all $p \in [1,\infty)$,*

$$W_p(\mu,\nu) = \left(\inf_{T_{\mu,\nu}\in\mathcal{M}} \int_{\mathbb{R}^n} |x - T_{\mu,\nu}(x)|^p d\mu(x)\right)^{\frac{1}{p}}, \quad \mu,\nu \in \mathscr{P}_p(X). \tag{39}$$

$\mathcal{M}$ is the set of all maps that rearrange the distribution μ into ν.

4.2 1D Problem

In [141], we proposed two ways of using W_2 in FWI were proposed. One can either compute the misfit globally by solving a 2D or 3D optimal transport problem or compare data trace-by-trace with the 1D explicit formula, see Theorem 1 below. For the 1D approach, the corresponding misfit function in FWI becomes

$$J_1(m) = \sum_{r=1}^{R} W_2^2(f(\mathbf{x_r},t;m), g(\mathbf{x_r},t)), \tag{40}$$

where R is the total number of time history traces, g is the observed data, f is the simulated data, $\mathbf{x_r}$ are the receiver locations, and m is the model parameter. Mathematically it is W_2 metric in the time domain and L^2 norm in the spatial domain.

For f and g in one dimension, it is possible to exactly solve the optimal transportation problem [131] in terms of the cumulative distribution functions

$$F(x) = \int_{-\infty}^{x} f(t)\,dt, \quad G(y) = \int_{-\infty}^{y} g(t)\,dt. \tag{41}$$

In fact, the optimal map is just the unique monotone rearrangement of the density f into g. In order to compute the Wasserstein metric (W_p), we need the cumulative distribution functions F and G and their inverses F^{-1} and G^{-1} as the following theorem states:

Theorem 1 (Optimal transportation on $\mathbb{R}$). *Let $0 < f,g < \infty$ be two probability density functions, each supported on a connected subset of $\mathbb{R}$. Then the optimal map from f to g is $T = G^{-1} \circ F$.*

From the theorem above, we derive another formulation for the 1D quadratic Wasserstein metric:

$$\begin{aligned} W_2^2(f,g) &= \int_0^1 |F^{-1} - G^{-1}|^2 dy \\ &= \int_X |x - G^{-1}(F(x))|^2 f(x) dx. \end{aligned} \tag{42}$$

The corresponding Fréchet derive which is also the adjoint source term in the backward propagation is:

$$\frac{\partial W_2^2(f,g)}{\partial f} = \left(\int_t^{T_0} -2\big(s - G^{-1}(F(s))\big) \frac{dG^{-1}(y)}{dy}\Big|_{y=F(s)} f(s)ds \right) dt + |t - G^{-1}(F(t))|^2 dt. \tag{43}$$

This adjoint source term in the discrete 1D setting can be computed as

$$\left[U \ \mathrm{diag}\left(\frac{-2f(t)dt}{g(G^{-1}\circ F(t))} \right) \right] (t - G^{-1}\circ F(t))dt + |t - G^{-1}\circ F(t)|^2 dt, \tag{44}$$

where U is the upper triangular matrix whose non-zero components are 1.

4.3 Monge-Ampère Equation

This fully nonlinear partial differential equation plays an important role in computing the Wasserstein metric.

4.3.1 Introduction

In the previous section, we introduced the 1D optimal transport technique of comparing seismic data trace by trace and the explicit solution formula. Another option is a general optimal transport problem in all dimensions. In the global case we compare the full datasets and consider the whole synthetic data f and observed data g as objects with the general quadratic Wasserstein metric (W_2):

$$J_2(m) = W_2^2(f(\mathbf{x_r},t;m), g(\mathbf{x_r},t)). \tag{45}$$

The simple exact formula for 1D optimal transportation does not extend to optimal transportation in higher dimensions. Nevertheless, it can be computed by relying on two important properties of the optimal mapping $T(x)$: conservation of mass and cyclical monotonicity. From the definition of the problem, $T(x)$ maps f into g. If T is a sufficiently smooth map and $\det(\nabla T(x)) \neq 0$, the change of variables formula formally leads to the requirement

$$f(x) = g(T(x))\det(\nabla T(x)). \tag{46}$$

The optimal map takes on additional structure in the special case of the cost function (i.e., $c(x,y) = |x-y|^2$): it is cyclically monotone [16, 68].

Definition 3 (Cyclical monotonicity). *We say that $T : X \to Y$ is cyclically monotone if for any $m \in \mathbb{N}^+$, $x_i \in X$, $1 \le i \le m$,*

$$\sum_{i=1}^m |x_i - T(x_i)|^2 \le \sum_{i=1}^m |x_i - T(x_{i-1})|^2 \tag{47}$$

or equivalently

$$\sum_{i=1}^m \langle T(x_i), x_i - x_{i-1}\rangle \ge 0 \tag{48}$$

where $x_0 \equiv x_m$.

Additionally, a cyclically monotone mapping is formally equivalent to the gradient of a convex function [16, 68]. Making the substitution $T(x) = \nabla u(x)$ into the constraint (46) leads to the Monge-Ampère equation

$$\det(D^2 u(x)) = \frac{f(x)}{g(\nabla u(x))}, \quad u \text{ is convex.} \tag{49}$$

In order to compute the misfit between distributions f and g, we first compute the optimal map $T(x) = \nabla u(x)$ via the solution of this Monge-Ampère equation coupled to the non-homogeneous Neumann boundary condition

$$\nabla u(x)\cdot \nu = x \cdot \nu, \ x \in \partial X. \tag{50}$$

The squared Wasserstein metric is then given by

$$W_2^2(f,g) = \int_X f(x)\,|x - \nabla u(x)|^2\, dx. \tag{51}$$

For the general Monge-Ampère equation, the uniqueness of the optimal map is not guaranteed. One need to discuss it in the context of a particular cost function and certain hypothesis. For example, the cyclical monotonicity is the key element in the proof of the following Brenier's theorem [16, 45] which gives an elegant result about the uniqueness of optimal transport map for the quadratic cost $|x-y|^2$:

Theorem 2 (Brenier's theorem). *Let μ and ν be two compactly supported probability measures on $\mathbb{R}^n$. If μ is absolutely continuous with respect to the Lebesgue measure, then*

1. *There is a unique optimal map T for the cost function $c(x,y) = |x-y|^2$.*
2. *There is a convex function $u : \mathbb{R}^n \to \mathbb{R}$ such that the optimal map T is given by $T(x) = \nabla u(x)$ for μ-a.e. x.*

Furthermore, if $\mu(dx) = f(x)dx$, $\nu(dy) = g(y)dy$, then T is differential μ-a.e. and

$$\det(\nabla T(x)) = \frac{f(x)}{g(T(x))}. \tag{52}$$

We are here considering the connection between the Monge-Ampère equation and optimal transport where the transport map is geometric in nature. The

Monge-Ampère equation is of course also known for many other connections to geometry and mathematical physics. Let us mention a few examples. It arises naturally in many problems such as affine geometry [32], Riemannian geometry [1], isometric embedding [60], reflector shape design [135], etc. In the last century, treatments about this equation mostly came from the geometric problems above [21, 31, 72, 90, 126]. If we consider the following general Monge-Ampère equation:

$$\det(D^2u(x)) = f(x,u,Du), \tag{53}$$

when $f = K(x)(1+|Du|^2)^{(n+2)/2}$, the equation becomes the prescribed Gaussian curvature equation [45]. In affine geometry, an affine sphere in the graph satisfies the Monge-Ampère equation (53). The affine maximal surface satisfies a fourth-order equation which is related to the general Monge-Ampère equation:

$$\sum_{i,j=i}^{n} U^{ij}\partial_{x_i}\partial_{x_j}\left[\det(D^2u)\right]^{-\frac{n+1}{n+2}} = 0, \tag{54}$$

where U^{ij} is the cofactor matrix of D^2u [127].

4.3.2 Weak Solutions

Although the Monge-Ampère equation is a second-order PDE, there is no guarantee that the classical C^2 solution always exists. For the generalized Monge-Ampère equation (53) with homogeneous Dirichlet boundary condition $u = 0$ on $\partial\Omega$, it is well-known that there exists a classical convex solution $u \in C^2(\Omega)\cup C(\overline{\Omega})$, when f is strictly positive and sufficiently smooth [22, 23, 24]. When the assumptions no longer hold, we solve for two types of weak solutions instead: the Aleksandrov solution and the viscosity solution. One can refer to [58] for more details and proofs of the following definitions and theorems.

Let Ω be the open subset of $\mathbb{R}^d$ and $u : \Omega \to \mathbb{R}$. We denote $\mathcal{P}(\mathbb{R}^d)$ as the set of all subsets of $\mathbb{R}^d$.

Definition 4. *The normal mapping of u, or the subdifferential of u, is the set-valued mapping $\partial u : \Omega \to \mathcal{P}(\mathbb{R}^d)$ defined by*

$$\partial u(x_0) = \{p : u(x) \geq u(x_0) + p\cdot(x-x_0),\ \text{for all } x \in \Omega\} \tag{55}$$

Given $V \in \Omega$, $\partial u(V) = \cup_{x\in V}\partial u(x)$.

Theorem 3 (Monge-Ampère measure). *If Ω is open and $u \in C(\Omega)$, then the class*

$$\mathcal{S} = \{V \subset \Omega : \partial u(V) \text{is Lebesgue measurable}\}$$

is a Borel σ-algebra. The set function $Mu : \mathcal{S} \to \overline{\mathbb{R}}$ defined by

$$Mu(V) = |\partial u(V)| \tag{56}$$

is a measure, finite on compact sets, called the Monge-Ampère measure associated with the function u.

This is a measure generated by the Monge-Ampère operator, which naturally defines the notion of the Aleksandrov solution.

Definition 5 (Aleksandrov solution). *Let ν be a Borel measure defined on Ω which is an open and convex subset of $\mathbb{R}^n$. The convex function u is a weak solution, in the sense of Aleksandrov, to the Monge-Ampère equation*

$$\det D^2u = \nu \quad \text{in } \Omega \tag{57}$$

if the associated Monge-Ampère measure Mu defined in (56) is equal to ν.

Next we state one existence and uniqueness result for the Aleksandrov solution [2].

Theorem 4 (Existence and uniqueness of the Aleksandrov solution). *Consider the following Dirichlet problem of the Monge-Ampère equation*

$$\begin{aligned} \det D^2u &= \nu \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{58}$$

on a convex bounded domain $\Omega \in \mathbb{R}^d$ with boundary $\partial\Omega$. Assume that ν is a finite Borel measure and $g \in C(\partial\Omega)$ which can be extended to a convex function $\tilde{g} \in C(\bar{\Omega})$. Then the Monge-Ampère equation (58) has a unique convex Aleksandrov solution in $C(\overline{\Omega})$.

Aleksandrov's generalized solution corresponds to the curvature measure in the theory of convex bodies [127]. A finite difference scheme for computing Aleksandrov measure induced by D^2u in 2D was conducted in [98] with the solution u comes as a byproduct [50].

Another notion of weak solution is the viscosity solution which occurs naturally if f is continuous in (53).

Definition 6 (Viscosity solution). *Let $u \in C(\Omega)$ be a convex function and $f \in C(\Omega)$, $f \geq 0$. The function u is a viscosity subsolution (supersolution) of (53) in Ω if whenever convex function $\phi \in C^2(\Omega)$ and $x_0 \in \Omega$ are such that $(u-\phi)(x) \leq (\geq)(u-\phi)(x_0)$ for all x in the neighborhood of x_0, then we must have*

$$\det(D^2\phi(x_0)) \leq (\geq) f(x_0).$$

The function u is a viscosity solution if it is both a viscosity subsolution and supersolution.

We can relate these two notions of weak solution in the following proposition:

Proposition 5. *If u is a Aleksandrov (generalized) solution of (53) with f continuous, then u is also a viscosity solution.*

4.4 Numerical Optimal Transport in Higher Dimensions

In this section, we will summarize some of the current numerical methods for solving the optimal transport problems in higher dimensions. These methods are based on the equivalent or relaxed formulations of the original Monge's problem. In the end, we will introduce a monotone finite difference Monge-Ampère solver which is proved to converge to the viscosity solution to (49) [4, 51].

4.4.1 General Methods

Optimal transport is a well-studied subject in mathematics while the computation techniques are comparatively underdeveloped. We will focus on analysis based methods. There are combinatorial techniques that typically are computationally costly in higher dimensions, for example, the Hungarian algorithm [73].

The definition (39) is the original static formulation of the optimal transport problem with a quadratic cost. It is an infinite dimensional optimization problem if we search for T directly. The non-symmetric nature of Monge's problem also generated difficulty because the map is unnecessarily bijective [76].

In the 40's, Kantorovich relaxed the constraints and formulated the dual problem [66]. Instead of searching for a map T, the transference plan γ is considered, which is also a measure supported by the product space $X \times Y$. The Kantorovich problem is the following:

$$\inf_{\gamma} \left\{ \int_{X\times Y} c(x,y)d\gamma \mid \gamma \geq 0 \text{ and } \gamma \in \Pi(\mu,\nu) \right\}, \tag{59}$$

where $\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(X\times Y) \mid (P_X)\sharp\gamma = \mu, (P_Y)\sharp\gamma = \nu\}$. Here (P_X) and (P_Y) denote the two projections, and $(P_X)\sharp\gamma$ and $(P_Y)\sharp\gamma$ are two measures obtained by pushing forward γ with these two projections.

Consider $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$, the Kantorovich dual problem is formulated as the following [131]:

$$\sup_{\varphi,\psi} \left(\int_X \varphi \, d\mu + \int_Y \psi \, d\nu \right), \tag{60}$$

subject to $\varphi(x) + \psi(y) \leq c(x,y)$, for any $(x,y) \in X \times Y$.

The dual formulation is a linear optimization problem which is solvable by linear programming [40, 97, 112]. Kantorovich obtained the 1975 Nobel prize in economics for his contributions to resource allocation problems where he interpreted the dual problem as an economic equilibrium. Recently Cuturi introduced the entropy regularized optimal transport problem which enforces the desirable properties for optimal transference plan and convexifies the problem. There have been extremely efficient computational algorithms [38] which allow various applications in image processing, neuroscience, machine learning, etc [9, 39, 54, 118].

In the 90's, Benamou and Brenier derived an equivalent dynamic formulation [7] which has been one of the main tools for numerical computation. The Benamou-Brenier formula identifies the squared quadratic Wasserstein metric between μ and ν by

$$W_2^2(\mu,\nu) = \inf \int_0^1 \int |v(t,x)|^2 \rho(t,x) dx dt, \tag{61}$$

where the infimum is taken among all the solutions of the continuity equation:

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla(v\rho) &= 0, \\ \text{subject to} \quad \rho(0,x) = f,\ \rho(1,x) &= g, \end{aligned} \tag{62}$$

In fact the infimum is taken among all Borel fields $v(t,x)$ that transports μ to ν continuously in time, satisfying the zero flux condition on the boundary. Many fast solvers based on this dynamic formulation has been proposed in literature [8, 77, 99]. They are used particularly in image registration, warping, texture mixing, etc.

4.4.2 The Finite Difference Monge-Ampère Solver

As we have seen for the quadratic Wasserstein distance, the optimal map can be computed via the solution of a Monge-Ampère partial differential equation [10]. This approach has the advantage of drawing on the well-developed field of numerical partial differential equations (PDEs). We solve the Monge-Ampère equation numerically for the viscosity solution using an almost-monotone finite difference method relying on the following reformulation of the Monge-Ampère operator, which automatically enforces the convexity constraint [51]. The scientific reason for using monotone type schemes follows from the following theorem by Barles and Souganidis [4]:

Theorem 6 (Convergence of Approximation Schemes [4]). *Any consistent, stable, monotone approximation scheme to the solution of fully nonlinear second-order elliptic or parabolic PDE converges uniformly on compact subsets to the unique viscosity solution of the limiting equation, provided this equation satisfies a comparison principle.*

The numerical scheme of [10] uses the theory of [4] to construct a convergent discretization of the Monge-Ampère equation (49) as stated in Theorem 7. A variational characterization of the determinant on the left hand side which also involves the negative

part of the eigenvalues was proposed as the following equation:

$$\det(D^2u) = \min_{\{v_1,v_2\}\in V} \{\max\{u_{v_1,v_1},0\}\max\{u_{v_2,v_2},0\} + \min\{u_{v_1,v_1},0\} + \min\{u_{v_2,v_2},0\}\} \tag{63}$$

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

Equation (63) can be discretized by computing the minimum over finitely many directions $\{v_1, v_2\}$, which may require the use of a wide stencil. In the low-order version of the scheme, the minimum in (63) is approximated using only two possible values. The first uses directions aligning with the grid axes.

$$MA_1[u] = \max\{\mathcal{D}_{x_1x_1}u,\delta\}\max\{\mathcal{D}_{x_2x_2}u,\delta\} + \min\{\mathcal{D}_{x_1x_1}u,\delta\} + \min\{\mathcal{D}_{x_2x_2}u,\delta\} - f/g(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u) - u_0. \tag{64}$$

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

$$MA_2[u] = \max\{\mathcal{D}_{vv}u,\delta\}\max\{\mathcal{D}_{v^\perp v^\perp}u,\delta\} + \min\{\mathcal{D}_{vv}u,\delta\} + \min\{\mathcal{D}_{v^\perp v^\perp}u,\delta\} - f/g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_vu+\mathcal{D}_{v^\perp}u),\frac{1}{\sqrt{2}}(\mathcal{D}_vu-\mathcal{D}_{v^\perp}u)\right) - u_0. \tag{65}$$

Then the monotone approximation of the Monge-Ampère equation is

$$M_M[u] \equiv -\min\{MA_1[u], MA_2[u]\} = 0. \tag{66}$$

We also define a second-order approximation, obtained from a standard centred difference discretization,

$$M_N[u] \equiv -\left((\mathcal{D}_{x_1x_1}u)(\mathcal{D}_{x_2x_2}u) - (\mathcal{D}_{x_1x_2}u^2)\right) + f/g(\mathcal{D}_{x_1}u,\mathcal{D}_{x_2}u) + u_0 = 0. \tag{67}$$

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

$$M_F[u] \equiv M_M[u] + \epsilon S\left(\frac{M_N[u]-M_M[u]}{\epsilon}\right) \tag{68}$$

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

$$S(x) = \begin{cases} x & |x| \le 1 \\ 0 & |x| \ge 2 \\ -x+2 & 1 \le x \le 2 \\ -x-2 & -2 \le x \le -1. \end{cases} \tag{69}$$

The Neumann boundary condition is implemented using standard one-sided differences. As described in [48, 51], the (formal) Jacobian $\nabla M_F[u]$ of the scheme can be obtained exactly. It is known to be sparse and diagonally dominant.

Theorem 7 (Convergence to Viscosity Solution [51, Theorem 4.4]). *Let the Monge-Ampère equation (49) have a unique viscosity solution and let $g > 0$ be Lipschitz continuous on $\mathbb{R}^d$. Then the solutions of the scheme (68) converge to the viscosity solution of (49) with a formal discretization error of $\mathcal{O}(Lh^2)$ where L is the Lipschitz constant of g and h is the resolution of the grid.*

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

$$W_2^2(f,g) \approx \sum_{j=1}^{n}(x_j - D_{x_j}u_h)^T \mathrm{diag}(f)(x_j - D_{x_j}u_h)dt, \tag{70}$$

where n is the dimension of the data f and g. Then the gradient of the discrete squared Wasserstein metric can be expressed as

$$\frac{\partial W_2^2(f,g)}{\partial f} = \sum_{j=1}^{n}\left[-2\nabla M_F^{-1}[u_f]^T D_{x_j}^T \mathrm{diag}(f)\right](x_j - D_{x_j}u_f)dt + \sum_{j=1}^{n}|x_j - D_{x_j}u_f|^2 dt, \tag{71}$$

This term is the discretized version of the Fréchet derivative of the misfit function (45) with respect to the synthetic data f, i.e., the adjoint source $\frac{\partial J}{\partial f}$ in the adjoint wave equation (36).

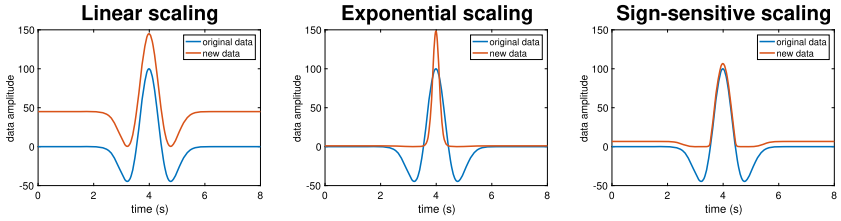
5. Application of Optimal Transport to Seismic Inversion

In this section, we first review the good properties of the W_2 norm for the application of full-waveform inversion. We will also explain some details of the implementations and show numerical results of using optimal transport based metrics as the misfit function in FWI.

5.1 W_2 Properties

As we demonstrated in [48], the squared Wasserstein metric has several properties that make it attractive as a choice for misfit function. One highly desirable feature is its convexity with respect to several parameterizations that occur naturally in seismic waveform inversion [141]. For example, variations in the wave velocity lead to simulated f that are derived from shifts,

$$f(x;s) = g(x+s\eta), \quad \eta \in \mathbb{R}^n, \tag{72}$$



(a) Linear (b) Exponential (c) Sign-sensitive

Figure 7. (a) The linear, (b) the exponential and (c) the sign-sensitive scaling of a Ricker wavelet (Blue).

or dilations,

$$f(x;A) = g(Ax), \quad A^T = A, A > 0, \tag{73}$$

applied to the observation g. Variations in the strength of a reflecting surface or the focusing of seismic waves can also lead to local rescalings of the form

$$f(x;\beta) = \begin{cases} \beta g(x), & x \in E \\ g(x), & x \in \mathbb{R}^n \backslash E. \end{cases} \tag{74}$$

Theorem 8 (Convexity of squared Wasserstein metric [48]). *The squared Wasserstein metric $W_2^2(f(m),g)$ is convex with respect to the model parameters m corresponding to a shift s in (72), the eigenvalues of a dilation matrix A in (73), or the local rescaling parameter β in (74).*

Another important property of optimal transport is the insensitivity to noise. All seismic data contains either natural or experimental equipment noise. For example, the ocean waves lead to extremely low-frequency data in the marine acquisition. Wind and cable motions also generate random noise.

Theorem 9 (Insensitivity to noise [48]). *Let f_{ns} be f with a piecewise constant additive noise of mean zero uniform distribution. The squared Wasserstein metric $W_2^2(f,f_{ns})$ is of $\mathcal{O}(\frac{1}{N})$ where N is the number of pieces of the additive noise in f_{ns}.*

The L^2 norm is known to be sensitive to noise since the misfit between clean and noisy data is calculated as the sum of squared noise amplitude at each sampling point.

5.2 Data Normalization

In optimal transport theory, there are two main requirements for signals f and g: positivity and mass balance. Since these are not expected for seismic signals, some data pre-processing is needed before we can implement Wasserstein-based FWI. In [47, 48], the signals were separated into positive and negative parts $f^+ = \max\{f,0\}$, $f^- = \max\{-f,0\}$ and scaled by the total mass $\langle f \rangle = \int_X f(x)\,dx$. Inversion was accomplished using the modified misfit function

$$W_2^2\left(\frac{f^+}{\langle f^+\rangle}, \frac{g^+}{\langle g^+\rangle}\right) + W_2^2\left(\frac{f^-}{\langle f^-\rangle}, \frac{g^-}{\langle g^-\rangle}\right). \tag{75}$$

While this approach preserves the desirable theoretical properties of convexity to shifts and noise insensitivity, it is not easy to combine with the adjoint-state method and more realistic examples. We require the scaling function to be differentiable so that it is easy to apply the chain rule when calculating the Fréchet derivative for FWI backpropagation and also better suited for the Monge-Ampère and the wave equation solvers.

There are other ways to rescale the datasets so that they become positive. For example, we can square the data as $\tilde{f} = f^2$ or extract the envelope of the data. These methods preserve the convexity concerning simple shifts, but we have lost the uniqueness: $f^2 = g^2$ does not imply $f = g$. As a result, more local minima are present since the fact that the misfit $J(f^2,g^2)$ is decreasing does not necessarily indicate that f is approaching g, not to mention the non-unique issue of the inverse problem itself.

Typically, we first scale the data f to be positive as $\tilde{f}$ and then normalize to ensure mass balance as $\tilde{f}/<\tilde{f}>$. We now introduce three normalization methods that are robust in realistic large-scale inversions: the linear scaling [141] (Figure 7a)

$$\tilde{f} = f + c_1, \quad c_1 \geq \max\{-f,-g\}, \tag{76}$$

the exponential scaling [104] (Figure 7b)

$$\tilde{f} = \exp(c_2 f), \quad c_2 > 0, \tag{77}$$

and the sign-sensitive scaling (Figure 7c)

$$\tilde{f} = \begin{cases} f + \frac{1}{c_3}, & f \geq 0 \\ \frac{1}{c_3}\exp(c_3 f), & f < 0 \end{cases}, \quad c_3 > 0. \tag{78}$$

If c_2 in (77) and c_3 in (78) are large enough, these two scaling methods keep the convexity of W_2 norm regarding simple shifts as shown in Figure 5c. From Taylor expansion, we can see that the scalings are

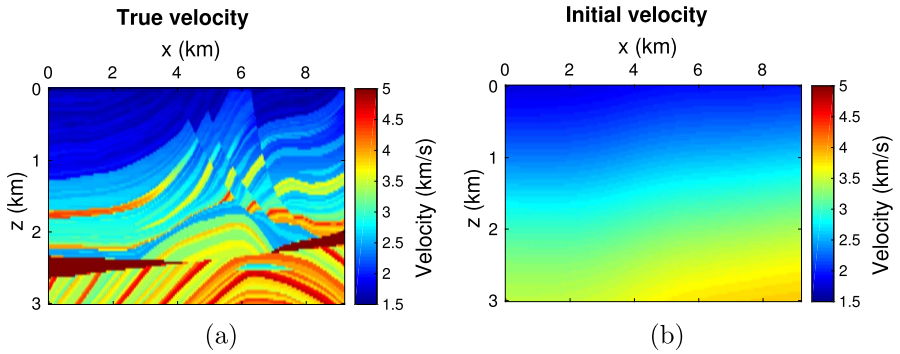


Figure 8. (a) True velocity and (b) inital velocity for full Marmousi model.

very close to the linear scaling when c_2 is small. One has to be careful with the exponential scaling (77) since it can easily become extremely large, but the sign-sensitive scaling (78) will not.

5.3 FWI with Kantorovich-Rubinstein Norm

When the cost function $c(x,y)$ is the L^1 norm $|x-y|$, i.e. $p=1$ in (39) with $f \geq 0$, $g \geq 0$, and $\int f = \int g$, the corresponding alternative W_1 distance has the following equivalent dual formulation:

$$W_1(f,g) = \max_{\varphi \in \mathrm{Lip}_1} \int_X \varphi(x)(f(x)-g(x))dx, \tag{79}$$

where Lip_1 is the space of Lipschitz continuous functions with Lipschitz constant 1. However, seismic data f and g are oscillatory containing both positive and negative parts. If $\int f \neq \int g$, the value of (79) is always $+\infty$. Recently, [88, 89] introduced the following Kantorovich-Rubinstein (KR) norm in FWI which is a relaxation of the original W_1 distance by constraining the dual space:

$$\mathrm{KR}(f,g) = \max_{\varphi \in \mathrm{BLip}_1} \int_X \varphi(x)(f(x)-g(x))dx \tag{80}$$

Here BLip_1 is the space of bounded Lipschitz continuous functions with Lipschitz constant 1. One advantage of using KR norm in FWI is that there is no need to normalize the data to be positive and mass balanced. However, KR norm has no direct connection with optimal transport once we no longer require f and g to be probability measures [130]. When f and g are far apart which is very common when the initial velocity is rough, the maximum in (80) is achieved by "moving" f^+ to f^- and g^+ to g^-. The notion of transport is void in this case and convexity is lost.

5.4 Numerical Results of Global W_2

In the next two subsections, we provide numerical results for two approaches to using W_2 with linear normalization (76): trace-by-trace comparison and using the entire 2D datasets as objects. Here a trace is the time history measured at one receiver while the entire dataset consists of the time history of all the receivers. These are compared with results produced by using the standard least-squares norm L^2 to measure the misfit. More examples can be found in [141].

First, we use a scaled Marmousi model to compare the inversion between global W_2 and the conventional L^2 misfit function. Figure 8a is the P-wave velocity of the true Marmousi model, but in this experiment, we use a scaled model which is 1 km in depth and 3 km in width. The inversion starts from an initial model that is the true velocity smoothed by a Gaussian filter with a deviation of 40, which is highly smoothed and far from the true model (a scaled version of Figure 8b). We place 11 evenly spaced sources on top at 50 m depth and 307 receivers on top at the same depth with a 10 m fixed acquisition. The discretization of the forward wave equation is 10 m in the x and z directions and 10 ms in time. The source is a Ricker wavelet which is the second derivative of the Gaussian function with a peak frequency of 15 Hz, and a bandpass filter is applied to remove the frequency components from 0 to 2 Hz.

We use L-BFGS, a quasi-Newton method as the optimization algorithm [78]. Inversions are terminated after 200 iterations. Figure 9a shows the inversion result using the traditional L^2 least-squares method after 200 L-BFGS iterations. The inversion result of global W_2 (Figure 9b) avoids the problem of local minima suffered by the conventional L^2 metric, whose result demonstrates spurious high-frequency artifacts due to a point-by-point comparison of amplitude.

We solve the Monge-Ampère equation numerically in each iteration of the inversion. The drawback to the PDE approach is that data must be sufficiently regular for solutions to be well-defined and for the numerical approximation to be accurate. To remain robust on realistic examples, we use filters that effectively smooth the seismic data, which can lead to a

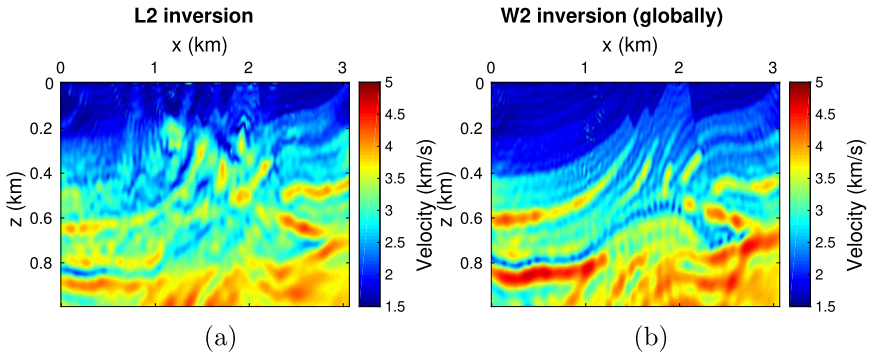


Figure 9. Inversion results of (a) L^2 and (b) global W_2 for the scaled Marmousi model.

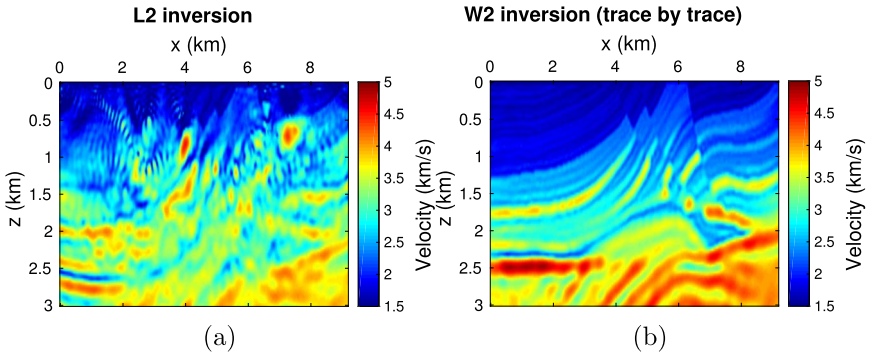


Figure 10. Inversion results of (a) L^2 and (b) trace-by-trace W_2 for the true Marmousi model.

loss of high-frequency information. For illustration in this paper, we perform computations using a Monge-Ampère solver for synthetic examples. Even in 2D, some limitations are apparent. This is expected to become even more of a problem in higher-dimensions and motivates our introduction of a trace-by-trace technique that relies on the exact 1D solution. The trace-by-trace technique is currently more promising for practical applications, as is evidenced in our computational examples in the next section.

5.5 Numerical Results of Trace-by-Trace W_2

Recall that for the 1D trace-by-trace approach, the misfit function in FWI is

$$J_1(m) = \sum_{r=1}^{R} W_2^2(f(\mathbf{x_r},t;m), g(\mathbf{x_r},t)), \tag{81}$$

where R is the total number of traces, g is observed data, f is simulated data, $\mathbf{x_r}$ are receiver locations, and m is the model parameter. The adjoint source term for each single trace is

$$\frac{\partial W_2^2(f,g)}{\partial f} = \left(\int_t^{T_0} \frac{-2(s - G^{-1} \circ F(s))}{g(G^{-1} \circ F(s))} f(s) ds + |t - G^{-1}(F(t))|^2 \right) dt. \tag{82}$$

The next experiment is to invert the full Marmousi model by conventional L^2 and trace-by-trace W_2 misfit. Figure 8a is the P-wave velocity of the true Marmousi model, which is 3 km in depth and 9 km in width. The inversion starts from an initial model that is the true velocity smoothed by a Gaussian filter with a deviation of 40 (Figure 8b). The rest of the settings are the same as the previous section. Inversions are terminated after 300 L-BFGS iterations. Figure 10a shows the inversion result using the traditional L^2 least-squares method and figure 10b shows the final result using trace-by-trace W_2 misfit function. Again, the result of L^2 metric has spurious high-frequency artifacts while W_2 correctly inverts most details in the true model. The convergence curves in Figure 11 show that W_2 reduces the relative misfit to 0.1 in 20 iterations while L^2 converges slowly to a local minimum.

5.6 Insensitivity to Noise

One of the good properties of the quadratic Wasserstein metric is the insensitivity to noise [48]. We repeat the previous experiment with a noisy reference by adding a uniform random iid noise to the data from the true velocity (Figure 12a). The signal-to-noise ratio (SNR) is -3.47 dB. In optimal transport,

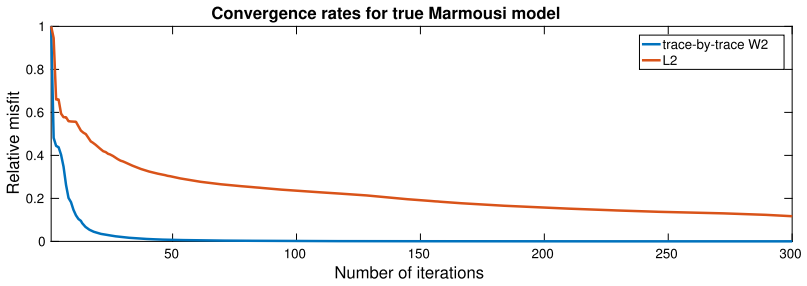


Figure 11. The convergence curves for trace-by-trace W_2 and L^2 based inversion of the full Marmousi model.

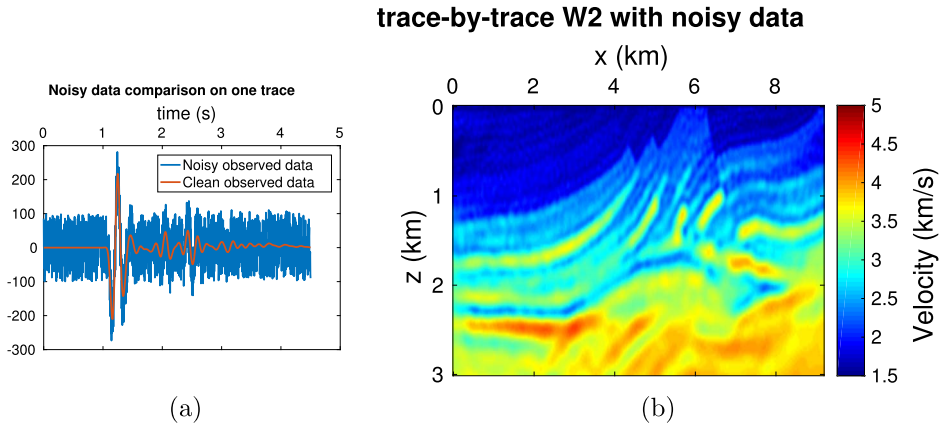


Figure 12. (a) Noisy and clean data and (b) inversion result with the noisy data.

the effect of noise is essentially negligible due to the strong cancellation between the nearby positive and negative noise.

All the settings remain the same as in the previous numerical experiment except the observed data. After 96 iterations, the optimization converges to a velocity presented in Figure 12b. Although the result has lower resolution than Figure 10b, it still recovers most features of Marmousi model correctly. Even when the noise is much larger than the signal, the quadratic Wasserstein metric still converges reasonably well.

References

[1] Thierry Aubin. *Some nonlinear problems in Riemannian geometry*. Springer Science & Business Media, 2013.

[2] Gerard Awanou. Discrete Aleksandrov solutions of the Monge-Ampère equation. *arXiv preprint arXiv:1408.1729*, 2014.

[3] Hyoungsu Baek, Henri Calandra, and Laurent Demanet. Velocity estimation via registration-guided least-squares inversion. *Geophysics*, 2014.

[4] Guy Barles and Panagiotis E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic analysis*, 4(3):271–283, 1991.

[5] Vladimir Bashkardin, Sergey Fomel, Parvaneh Karimi, Alexander Klokov, and Xiaolei Song. Sigsbee model.

[6] Edip Baysal, Dan D. Kosloff, and John W. C. Sherwood. Reverse time migration. *Geophysics*, 48(11):1514–1524, Nov 1983.

[7] 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.

[8] Jean-David Benamou and Guillaume Carlier. Augmented lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1):1–26, 2015.

[9] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[10] Jean-David Benamou, Brittany D. Froese, and Adam M. Oberman. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.

[11] Wafik B. Beydoun and Albert Tarantola. First Born and Rytov approximations: Modeling and inversion conditions in a canonical example. *The Journal of the Acoustical Society of America*, 83(3):1045–1055, Mar 1988.

[12] Harmen Bijwaard, Wim Spakman, and E. Robert Engdahl. Closing the gap between regional and global travel time tomography. *Journal of Geophysical Research: Solid Earth*, 103(B12):30055–30078, Dec 1998.

[13] Biondo Biondi and Ali Almomin. Tomographic full waveform inversion (TFWI) by combining full waveform inversion with wave-equation migration velocity analysis. In *SEG Technical Program Expanded Abstracts 2012*, pages 1–5. Society of Exploration Geophysicists, 2012.

[14] Michel Bouchon and Francisco J. Sánchez-Sesma. Boundary integral equations and boundary elements methods in elastodynamics. *Advances in geophysics*, 48:157–189, 2007.

[15] Ebru Bozdağ, Jeannot Trampert, and Jeroen Tromp. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870, May 2011.

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

[17] Romain Brossier, Stéphane Operto, and Jean Virieux. Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *Geophysics*, 74(6):WCC105–WCC118, 2009.

[18] Romain Brossier, Stéphane Operto, and Jean Virieux. Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3):R37–R46, 2010.

[19] Kenneth P. Bube and Robert T. Langan. Hybrid L^1/L^2 minimization with applications to tomography. *Geophysics*, 62(4):1183–1195, Jul 1997.

[20] Carey Bunks, Fatimetou M. Saleck, S. Zaleski, and G. Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, 1995.

[21] L. Caffarelli, J. J. Kohn, L. Nirenberg, and J. Spruck. The Dirichlet problem for nonlinear second-order elliptic equations. II. Complex Monge-Ampère, and uniformaly elliptic, equations. *Communications on Pure and Applied Mathematics*, 38(2):209–252, Mar 1985.

[22] Luis A. Caffarelli. Interior a Priori Estimates for Solutions of Fully Non-Linear Equations. *The Annals of Mathematics*, 130(1):189, Jul 1989.

[23] Luis A. Caffarelli. Interior $W^{2,p}$ Estimates for Solutions of the Monge-Ampère Equation. *The Annals of Mathematics*, 131(1):135, Jan 1990.

[24] Luis A. Caffarelli. Some regularity properties of solutions of Monge Ampère equation. *Communications on Pure and Applied Mathematics*, 44(8-9):965–969, Oct 1991.

[25] Emmanuel Candès, Laurent Demanet, David Donoho, and Lexing Ying. Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899, 2006.

[26] V. Červený, I. A. Molotkov, and I. Pšenčík. Ray method in seismology: Charles univ, 1977.

[27] Vlastislav Červený, Mikhail M. Popov, and Ivan Pšenčík. Computation of wave fields in inhomogeneous media—Gaussian beam approach. *Geophysical Journal International*, 70(1):109–128, 1982.

[28] Vlastislav Červený and Ivan Pšenčík. Seismic, ray theory. pages 1244–1258. Springer, Dordrecht, 2011.

[29] Guiting Chen and Zhenli Wang. Robust full-waveform inversion based on particle swarm optimization. In *SEG Technical Program Expanded Abstracts 2017*, pages 1302–1306. Society of Exploration Geophysicists, Aug 2017.

[30] Jing Chen, Yifan Chen, Hao Wu, and Dinghui Yang. The quadratic Wasserstein metric for earthquake location. *arXiv preprint arXiv:1710.10447*, 2017.

[31] Shiu-Yuen Cheng and Shing-Tung Yau. On the regularity of the Monge-Ampère equation $det(\partial^2 u/\partial x_i \partial x_j) = f(x,u)$. *Communications on Pure and Applied Mathematics*, 30(1):41–68, Jan 1977.

[32] Shiu-Yuen Cheng and Shing-Tung Yau. Complete affine hypersurfaces. Part I. The completeness of affine metrics. *Communications on Pure and Applied Mathematics*, 39(6):839–866, Nov 1986.

[33] Jon F. Claerbout. Toward a unified theory of reflector mapping. *Geophysics*, 36(3):467–481, Jun 1971.

[34] Jon F. Claerbout. *Imaging the earth's interior*. Blackwell scientific publications Oxford, 1985.

[35] Jon F. Claerbout. *Earth soundings analysis: Processing versus inversion*, volume 6. Blackwell Scientific Publications London, 1992.

[36] Robert Clayton and Björn Engquist. Absorbing boundary conditions for acoustic and elastic wave equations. *Bulletin of the Seismological Society of America*, 67(6):1529–1540, 1977.

[37] E. Crase, A. Pica, M. Noble, J. McDonald, and A. Tarantola. Robust elastic nonlinear waveform inversion: Application to real data. *Geophysics*, 55(5):527–538, May 1990.

[38] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[39] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[40] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational Wasserstein problems. *arXiv preprint arXiv:1503.02533*, 2015.

[41] F. A. Dahlen and Adam M. Baig. Fréchet kernels for body-wave amplitudes. *Geophysical Journal International*, 150(2):440–466, Aug 2002.

[42] Wei Dai, Hao Deng, Kun Jiao, and Denes Vigh. Least-squares reverse-time migration: An example of subsalt imaging. In *SEG Technical Program Expanded Abstracts 2017*, pages 4427–4431. Society of Exploration Geophysicists, Aug 2017.

[43] Wei Dai, Paul Fowler, and Gerard T. Schuster. Multi-source least-squares reverse time migration. *Geophysical Prospecting*, 60(4):681–695, Jul 2012.

[44] Wei Dai and Gerard T. Schuster. Plane-wave least-squares reverse-time migration. *Geophysics*, 78(4):S165–S177, Jul 2013.

[45] Guido De Philippis and Alessio Figalli. The Monge-Ampère equation and its link to optimal transportation. Oct 2013.

[46] Laurent Demanet. Waves and imaging class notes – 18.325. 2016.

[47] B. Engquist and B. D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5):979–988, 2014.

[48] Bjorn Engquist, Brittany D. Froese, and Yunan Yang. Optimal transport for seismic full waveform inversion. *Communications in Mathematical Sciences*, 14(8):2309–2330, 2016.

[49] John Etgen, Samuel H. Gray, and Yu Zhang. An overview of depth imaging in exploration geophysics. *Geophysics*, 74(6):WCA5–WCA17, Nov 2009.

[50] Xiaobing Feng and Michael Neilan. Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation. *Journal of Scientific Computing*, 47(3):303–327, 2011.

[51] 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.

[52] Hiroyuki Fujiwara. The fast multipole method for solving integral equations of three-dimensional topography and basin problems. *Geophysical Journal International*, 140(1):198–210, 2000.

[53] Takashi Furumura, B. L. N. Kennett, and Hiroshi Takenaka. Parallel 3-D pseudospectral simulation of seismic wave propagation. *Geophysics*, 63(1):279–288, 1998.

[54] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.
[55] Samuel H. Gray. Seismic, migration. pages 1236–1244. Springer, Dordrecht, 2011.
[56] Antoine Guitton. Amplitude and kinematic corrections of migrated images for nonunitary imaging operators. *Geophysics*, 69(4):1017–1024, 2004.
[57] Antoine Guitton and William W. Symes. Robust inversion of seismic data using the Huber norm. *Geophysics*, 68(4):1310–1319, Jul 2003.
[58] Cristian E. Gutiérrez. *The Monge-Ampère Equation*, volume 89. Birkhäuser, 2016.
[59] Taeyoung Ha, Wookeen Chung, and Changsoo Shin. Waveform inversion using a back-propagation algorithm and a huber function norm. *Geophysics*, 74(3):R15–R24, 2009.
[60] Qing. Han and Jiaxing Hong. *Isometric embedding of Riemannian manifolds in Euclidean spaces*. American Mathematical Society, 2006.
[61] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, Jun 2013.
[62] Guanghui Huang, Rami Nammour, and William Symes. Full-waveform inversion via source-receiver extension. *Geophysics*, 82(3):R153–R171, 2017.
[63] Guanghui Huang, Huazhong Wang, and Haoran Ren. Two new gradient precondition schemes for full waveform inversion. *arXiv preprint arXiv:1406.1864*, 2014.
[64] J. A. Hudson and J. R. Heritage. The use of the Born approximation in seismic scattering problems. *Geophysical Journal International*, 66(1):221–240, 1981.
[65] Ursula Iturrarán-Viveros and Francisco J. Sánchez-Sesma. Seismic Wave Propagation in Real Media: Numerical Modeling Approaches. pages 1200–1210. Springer, Dordrecht, 2011.
[66] Leonid Vital'evich Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422, 1960.
[67] Martin Käser and Michael Dumbser. An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes—i. the two-dimensional isotropic case with external source terms. *Geophysical Journal International*, 166(2):855–877, 2006.
[68] M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.
[69] Pierre Kolb, Francis Collino, and Patrick Lailly. Prestack inversion of a 1-D medium. *Proceedings of the IEEE*, 74(3):498–508, 1986.
[70] Soheil Kolouri, Serim Park, Matthew Thorpe, Dejan Slepčev, and Gustavo K. Rohde. Transport-based analysis, modeling, and learning from signal and data distributions. *arXiv preprint arXiv:1609.04767*, 2016.
[71] Dimitri Komatitsch and Jeroen Tromp. Introduction to the spectral element method for three-dimensional seismic wave propagation. *Geophysical journal international*, 139(3):806–822, 1999.
[72] N. V. Krylov. On the general notion of fully nonlinear second-order elliptic equations. *Transactions of the American Mathematical Society*, 347(3):857–895, Mar 1995.
[73] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.
[74] 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.
[75] Patrick Lailly. Migration methods: partial but efficient solutions to the seismic inverse problem. *Inverse problems of acoustic and elastic waves*, 51:1387–1403, 1984.
[76] Bruno Lévy and Erica Schwindt. Notions of optimal transport theory and how to implement them on a computer. 2017.
[77] Wuchen Li, Ernest K. Ryu, Stanley Osher, Wotao Yin, and Wilfrid Gangbo. A parallel method for earth mover's distance. *UCLA Comput. Appl. Math. Pub. (CAM) Rep*, pages 17–12, 2017.
[78] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
[79] Jiangbo Liu, Hervé Chauris, and Henri Calandra. The normalized integration method-an alternative to full waveform inversion? In *25th Symposium on the Application of Geophysics to Engineering & Environmental Problems*, 2012.
[80] Zhenyue Liu and Norman Bleistein. Migration velocity analysis: Theory and an iterative algorithm. *Geophysics*, 60(1):142–153, 1995.
[81] Jianfeng Lu and Xu Yang. Frozen Gaussian approximation for high frequency wave propagation. *Communications in Mathematical Sciences*, 9(3):663–683, 2011.
[82] Jingrui Luo and Ru-Shan Wu. Seismic envelope inversion: Reduction of local minima and noise resistance. *Geophysical Prospecting*, 63(3):597–614, 2015.
[83] Simon Luo and Dave Hale. Least-squares migration in the presence of velocity errors. *Geophysics*, 79(4):S153–S161, 2014.
[84] Yi Luo and Gerard T. Schuster. Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653, 1991.
[85] Yong Ma and Dave Hale. Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion. *Geophysics*, 78(6):R223–R233, 2013.
[86] Kurt J. Marfurt. Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. *Geophysics*, 49(5):533–549, 1984.
[87] H. Marquering, F. A. Dahlen, and G. Nolet. Three-dimensional sensitivity kernels for finite-frequency traveltimes: the banana-doughnut paradox. *Geophysical Journal International*, 137(3):805–815, Jun 1999.
[88] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysical Journal International*, 205(1):345–377, 2016.
[89] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. An optimal transport approach for seismic tomography: application to 3D full waveform inversion. *Inverse Problems*, 32(11):115008, 2016.
[90] Hermann Minkowski. Volumen und Oberfläche. pages 146–192. Springer, Vienna, 1989.
[91] Peter Moczo, Johan O. A. Robertsson, and Leo Eisner. The finite-difference time-domain method for modeling of seismic wave propagation. *Advances in Geophysics*, 48:421–516, 2007.
[92] Gaspard Monge. Mémoire sur la théorie des déblais et de remblais. histoire de l'académie royale des sciences de paris. *avec les Mémoires de Mathématique et de Physique pour la même année*, pages 666–704, 1781.
[93] Peter Mora. Elastic wave-field inversion of reflection

and transmission data. *Geophysics*, 53(6):750–759, Jun 1988.

[94] Peter Mora. Inversion = migration + tomography. *Geophysics*, 54(12):1575–1586, Dec 1989.

[95] F. Natterer. An error bound for the Born approximation. *Inverse problems*, 20(2):447–452, 2004.

[96] Roger G. Newton. *Scattering theory of waves and particles*. Springer Science & Business Media, 2013.

[97] Adam M. Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

[98] V. I. Oliker and L. D. Prussner. On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ and its discretizations, i. *Numerische Mathematik*, 54(3):271–293, 1989.

[99] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[100] 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.

[101] Tim Pointer, Enru Liu, and John A. Hudson. Numerical modelling of seismic waves scattered by hydrofractures: application of the indirect boundary element method. *Geophysical Journal International*, 135(1):289–303, 1998.

[102] R. Gerhard Pratt. Inverse Theory Applied to Multi-Source Cross-Hole Tomography. Part 2: Elastic Wave-Equation. *Geophysical Prospecting*, 38(3):311–329, Apr 1990.

[103] R. Gerhard Pratt and M. H. Worthington. Inverse theory applied to multi-source cross-hole tomography. Part 1: Acoustic wave-equation method. *Geophysical prospecting*, 38(3):287–310, 1990.

[104] Lingyun Qiu, Jaime Ramos-Martínez, Alejandro Valenciano, Yunan Yang, and Björn Engquist. Full-waveform inversion with an exponentially encoded optimal-transport norm. In *SEG Technical Program Expanded Abstracts 2017*, pages 1286–1290. Society of Exploration Geophysicists, 2017.

[105] Yingming Qu, Zhenchun Li, Jianping Huang, and Jinli Li. Viscoacoustic anisotropic full waveform inversion. *Journal of Applied Geophysics*, 136:484–497, Jan 2017.

[106] N. Rawlinson and M. Sambridge. Seismic traveltime tomography of the crust and lithosphere. *Advances in Geophysics*, 46:81–199, 2003.

[107] Nicholas Rawlinson, Gregory Austin Houseman, and Clive D. N. Collins. Inversion of seismic refraction and wide-angle reflection traveltimes for three-dimensional layered crustal structure. *Geophysical Journal International*, 145(2):381–400, 2001.

[108] Kabir Roy Chowdhury. Seismic data acquisition and processing. pages 1081–1097. Springer, Dordrecht, 2011.

[109] Malcolm Sambridge and Kerry Gallagher. Inverse Theory, Monte Carlo Method. pages 639–644. Springer, Dordrecht, 2011.

[110] Paul Sava and Biondo Biondi. Wave-equation migration velocity analysis. I. Theory. *Geophysical Prospecting*, 52(6):593–606, 2004.

[111] H. Schiessel, R. Metzler, A. Blumen, and T. F. Nonnenmacher. Generalized viscoelastic models: their fractional equations with solutions. *Journal of physics A: Mathematical and General*, 28(23):6567, 1995.

[112] Bernhard Schmitzer. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016.

[113] Gerard T. Schuster. Seismic Imaging, Overview. pages 1121–1134. Springer, Dordrecht, 2011.

[114] Gji Seismology, Ebru Bozdağ, Jeannot Trampert, and Jeroen Tromp. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophys. J. Int*, 185:845–870, 2011.

[115] Mrinal K. Sen and Paul L. Stoffa. Inverse Theory, Global Optimization. pages 625–632. Springer, Dordrecht, 2011.

[116] Peter M. Shearer. *Introduction to seismology*. Cambridge University Press, 2009.

[117] Laurent Sirgue and R. Gerhard Pratt. Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies. *Geophysics*, 69(1):231–248, 2004.

[118] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[119] William W. Symes. Migration velocity analysis and waveform inversion. *Geophysical prospecting*, 56(6):765–790, 2008.

[120] W. W. Symes and James J. Carazzone. Velocity inversion by differential semblance optimization. *Geophysics*, 56(5):654–663, 1991.

[121] A. Tarantola. Inverse problems theory. *Methods for Data Fitting and Model Parameter Estimation. Elsevier, Southampton*, 1987.

[122] Albert Tarantola. Linearized inversion of seismic reflection data. *Geophysical prospecting*, 32(6):998–1015, 1984.

[123] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[124] Albert Tarantola and Bernard Valette. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics*, 20(2):219–232, 1982.

[125] Sven Treitel and Larry Lines. Past, present, and future of geophysical inversion—a new millennium analysis. *Geophysics*, 66(1):21–24, Jan 2001.

[126] Neil S. Trudinger and Xu-Jia Wang. The Monge-Ampère equation and its geometric applications. *Handbook of geometric analysis*, 1:467–524, 2008.

[127] Neil S. Trudinger and Xu-Jia Wang. The Monge-Ampère equation and its geometric applications. pages 467–524, 2008.

[128] Gunther Uhlmann. Travel time tomography. *J. Korean Math. Soc*, 38(4):711–722, 2001.

[129] Léon Van Hove. Correlations in space and time and Born approximation scattering in systems of interacting particles. *Physical Review*, 95(1):249, 1954.

[130] Anatoly Moiseevich Vershik. Long history of the Monge-Kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4):1–9, 2013.

[131] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[132] J. Virieux, A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou. 6. An introduction to full waveform inversion. In *Encyclopedia of Exploration Geophysics*, pages R1-1–R1-40. Society of Exploration Geophysicists, Jan 2014.

[133] Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

[134] Ping Wang, Adriano Gomes, Zhigang Zhang, and Ming Wang. Least-squares rtm: Reality and possibilities for

subsalt imaging. In *SEG Technical Program Expanded Abstracts 2016*, pages 4204–4209. Society of Exploration Geophysicists, Sep 2016.
[135] Xu-Jia Wang. On the design of a reflector antenna. *Inverse Problems*, 12(3):351–375, Jun 1996.
[136] Mike Warner and Lluís Guasch. Adaptive waveform inversion: Theory. In *SEG Technical Program Expanded Abstracts 2014*, pages 1089–1093. 2014.
[137] Arthur B. Weglein, Fernanda V. Araújo, Paulo M. Carvalho, Robert H. Stolt, Kenneth H. Matson, Richard T. Coates, Dennis Corrigan, Douglas J. Foster, Simon A. Shaw, and Haiyan Zhang. Inverse scattering series and seismic exploration. *Inverse problems*, 19(6):R27, 2003.
[138] Pengliang Yang, Romain Brossier, Ludovic Métivier, and Jean Virieux. A review on the systematic formulation of 3-D multiparameter full waveform inversion in viscoelastic medium. *Geophysical Journal International*, 207(1):129–149, 2016.
[139] Yunan Yang and Björn Engquist. Analysis of optimal transport related misfit functions in seismic imaging. In *International Conference on Geometric Science of Information*, pages 109–116. Springer, 2017.
[140] Yunan Yang and Björn Engquist. Analysis of optimal transport and related misfit functions in full-waveform inversion. *Geophysics*, 83(1):A7–A12, 2018.
[141] Yunan Yang, Björn Engquist, Junzhe Sun, and Brittany D. Froese. Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):1–103, 2017.
[142] Öz Yilmaz. *Seismic Data Analysis*. Society of Exploration Geophysicists, Jan 2001.
[143] Kwangjin Yoon, Kurt J. Marfurt, and William Starr. Challenges in reverse-time migration. In *SEG Technical Program Expanded Abstracts 2004*, pages 1057–1060. Society of Exploration Geophysicists, 2004.
[144] Kwangjin Yoon, Kurt J. Marfurt, and William Starr. Challenges in reverse-time migration. In *SEG Technical Program Expanded Abstracts 2004*, pages 1057–1060. Society of Exploration Geophysicists, Jan 2004.
[145] C. A. Zelt. Modelling strategies and model assessment for wide-angle seismic traveltime data. *Geophysical Journal International*, 139(1):183–204, Oct 1999.
[146] Colin A. Zelt. Traveltime tomography using controlled-source seismic data. pages 1453–1473. Springer, Dordrecht, 2011.
[147] Chong Zeng, Shuqian Dong, and Bin Wang. A guide to least-squares reverse time migration for subsalt imaging: Challenges and solutions. *Interpretation*, 5(3):SN1–SN11, Aug 2017.
[148] Hejun Zhu and Sergey Fomel. Building good starting models for full-waveform inversion using adaptive matching filtering misfit. *Geophysics*, 81(5):U61–U72, 2016.