A robust waveform inversion using a global comparison of modeled and observed data
Bingbing Sun1 and Tariq Alkhalifah†
https://doi.org/10.1190/tle38030185.1.

Abstract
A high-resolution model of the subsurface is the product of a successful full-waveform inversion (FWI) application. However, this optimization problem is highly nonlinear, and thus, we iteratively update the subsurface model by minimizing a misfit function that measures the difference between observed and modeled data. The L2-norm misfit function provides a simple, sample-by-sample comparison between the observed and modeled data. However, it is susceptible to local minima in the objective function if the low-wavenumber components of the initial model are not accurate enough. We review an alternative formulation of FWI based on a more global comparison. A combination of Radon transform and utilizing a matching filter allows for comparisons beyond sample to sample. We combine two recent developments to suggest the following algorithm for optimal inversion: (1) we compute the matching filter between the observed and modeled data in the Radon domain, which helps reduce the crosstalk introduced in the deconvolution step of computing the matching filter, and (2) we use Wasserstein distance to measure the distance between the resulting matching filter in the Radon domain and a representation of the Dirac delta function, which provides us with the optimal transport between the two distribution functions. We use a modified Marmousi model to show how this Radon-domain optimal-transport-based matching-filter approach can mitigate cycle skipping. Starting from a rather simplified \(v(z)\) media as the initial model, the proposed method can invert for the Marmousi model with considerable accuracy, while standard L2-norm formulation is trapped in a local minimum. Application of the proposed method to an offshore data set further demonstrates its robustness and effectiveness.

Introduction
Full-waveform inversion (FWI) (Virieux and Operto, 2009) aims to use the whole recorded waveform signal and all seismic events (diving waves, pre- and postcritical reflections, and converted waves) to invert for physical properties such as P- and S-wave velocities, attenuation, density, and anisotropy parameters. FWI normally yields high-resolution estimates of the medium compared to tomographic methods, up to the theoretical limit of half of the shortest wavelength of the recorded signal.

Designing a misfit function is an important component of FWI: a well-behaved misfit function would relax the requirement for a good initial velocity model or usable low-frequency signals in the data. In some circumstances, where the subsurface is complex or has large velocity contrasts, such as salt bodies (Wang et al., 2008), the low-resolution model output from tomography (Woodward et al., 2008) is insufficient and can hardly meet the requirement for FWI to avoid local minima when the inverted model may be far from the true model but still gives relatively small misfit values. This phenomenon is also referred to as cycle skipping in which the initial predicted data are shifted more than a half cycle compared to the measured data. Due to the low signal-to-noise ratio and acquisition limitations, obtaining ultra-low-frequency seismic signals tends to be difficult. A practical, cost-effective solution to cycle skipping in FWI is to design a misfit function that is reasonably convex and allows the model to gradually converge to the exact solution from a far-away, cycle-skipped initial model. Although the least-squares L2 norm was widely used as a misfit function for its simplicity and its potential for high-resolution models, it suffers from cycle skipping.

Recently, more advanced misfit functions were proposed, such as a matching-filter misfit function (Van Leeuwen and Mulder, 2008, 2010; Luo and Sava, 2011; Warner and Guasch, 2016) and the optimal-transport misfit function (Engquist and Froese, 2014; Engquist et al., 2016; Metivier et al., 2016; Qiu et al., 2017; Yang et al., 2018; Yang and Engquist, 2018). Those newly proposed methods transform the local sample-by-sample comparison to a more global trace-by-trace comparison. As expected, the resulting misfit function shows more convex behavior and can reasonably mitigate the local minima issue. The idea behind the matching-filter approach is that instead of comparing the seismic trace sample by sample through, for example, subtraction in the L2 norm, we first compute a matching filter by deconvolving the measured data from the modeled ones. If the velocity used in modeling is correct, the resulting matching filter will focus energy mainly at zero lag, providing a band-limited Dirac delta function. Otherwise, by penalizing the coefficients at a non-zero time lag, we can formulate an optimization problem. Luo and Sava (2011) propose a misfit function using deconvolution instead of crosscorrelation (Van Leeuwen and Mulder, 2008, 2010). Warner and Guasch (2016) introduce an extra normalization term. In their study, the normalization term plays an important role in providing better convexity for the misfit function and accelerating its convergence. This normalization term has its roots in the instantaneous travelt ime objective function (Alkhalifah and Choi, 2014) as demonstrated by Sun and Alkhalifah (2018b). As shown in Sun and Alkhalifah (2018a), the normalization term will occur naturally if we try to measure the Wasserstein distance between the resulting matching filter and the Dirac delta function. The normalization term is a strict requirement for the comparison of two distributions, especially to satisfy the mass conservation rule. As long as we include the normalization by preconditioning the resulting matching filter, we form an optimization that focuses the resulting matching filter to zero lag or near it. Although a
simple penalty term appears to try to focus the matching filter in the same way, it can also actually reduce the amplitude of the matching filter rather than push it to be a Dirac delta function.

All these matching-filter approaches extend the data comparison along the whole trace. However, while it is needed to avoid cycle skipping, the global comparison can introduce crosstalk between events in the measured and modeled data that are not related. Such crosstalk in the global comparison can cause drawbacks in the application. For example, Huang et al. (2017) report that the matching-filter approach deals mainly with diving waves and suffers in the presence of strong multiples. Thus, we propose introducing to the general FWI workflow a Radon transform applied to the data prior to computing of the matching filter. Radon transform is a mathematical technique that has been widely used in seismic data processing (Stoffa et al., 1981; Thorson and Claerbout, 1985). Here, we expand the application of Radon transform to FWI for designing a robust misfit function by integrating it with the matching-filter approach. For the Radon-domain matching-filter FWI, instead of computing the matching filter for each trace directly, we first transform a common-shot gather into the tau-\(p\) domain using Radon transform and then compute the matching filter for each trace indexed by the same slope \(p\)-value. Thus, instead of designing the matching filter by applying the matching on a single trace, we apply the matching on traces that have been stacked over slopes and represented in the tau-\(p\) domain; consequently, we intuitively take the coherency of the events into consideration. The criterion for a successful convergence is still the same in the tau-\(p\) domain; i.e., the energy of the matching filter focuses at zero lag.

Recently, optimal-transport-based misfit functions have demonstrated their potential. The idea of using the optimal-transport measure of the misfit for waveform inversion was first proposed by Engquist and Froese (2014). The Wasserstein metric is a concept based on optimal transportation (Villani, 2003). To use seismic data in optimal transport, the data sets of the modeled and measured seismic signals need to be modified and preconditioned to form density functions of probability distributions, which can be seen as the distribution of two piles of sand with equal mass. Different transportation plans of one collection into the other result in a different amount of energy spent in the transport. The plan with the lowest amount of energy needed in the transportation is the optimal map, and this lowest cost is defined as the Wasserstein metric. If we consider time signals, the resulting Wasserstein metric yields a convex function concerning time-shifted signals (Engquist and Froese, 2014). This property makes the Wasserstein metric a viable remedy to the limitations of the L2-norm misfit function in FWI applications.

The main issue in the implementation of optimal transport in FWI is the need to measure the distance between distributions. The density function for a probability distribution has two features: (1) its elements are nonnegative, and (2) it satisfies the mass conservation rule (its integration equals one). The nonnegative rule requirement is more critical as seismic signals are oscillatory with often zero mean; it hardly meets these two requirements for distribution.

There are currently two categories of optimal-transport-based misfit functions. The first category is related to the quadratic Wasserstein metric W2 (Engquist et al., 2016; Yang et al., 2018; Yang and Engquist, 2018). The W2 metric has more convex properties, and for a 1D problem, the explicit solution for the optimal map exists, which helps in solving the optimal-transport optimization problem (Villani, 2003). In this category, the seismic signal is modified to satisfy the requirements of a distribution. Yang and Engquist (2018) discuss different data modification strategies. Qiu et al. (2017) implement an exponentially encoded optimal-transport norm. Metivier et al. (2018) provide a review of current modification strategies. They show that the current application of such modifications to seismic signals needed to satisfy the distribution requirement has its drawbacks. For example, one strategy of positive and negative value separated comparison (Engquist and Froese, 2014), while it ensures convexity, is in violation of the mass conservation rule and is highly sensitive to the source phase rotation. Another strategy based on an affine scaling of the data (adding a relatively large value to make the signal positive) (Yang et al., 2018) will destroy the convexity with regard to time shifts.

Another category of optimal transport is related to using the Kantorovich-Rubinstein (KR) norm (Metivier et al., 2016). The KR norm is a relaxation of the Wasserstein distance, which is another optimal-transport metric with the absolute value as a cost function. The advantage of the KR norm is that it does not require the data to satisfy nonnegativity or mass conservation conditions. However, in this approach, unlike Wasserstein W2, there is no explicit solution for the optimal map, and an optimal-transport optimization problem has to be solved. This approach also does not maintain the convexity of the quadratic Wasserstein metric with respect to time shifts (Metivier et al., 2018).

Thus, we propose a framework for misfit-function designing by combining the matching-filter approach with the optimal-transport theory, resulting in a more elegant way to measure the distance between the matching filter and the Dirac delta function. Instead of applying the optimal transport between seismic signals, we apply it between the filter that matches the observed data to the modeled ones and the form we want this filter to be, which is a representation of the Dirac delta function. This matching filter naturally adheres to the requirements of the optimal-transport theory as we seek to transport the matching filter to a distribution given by the Dirac delta function. Since we use the W2 measure, unlike the dual optimal-transport strategy, we inherently obtain the explicit solution for the optimal map, which makes the computation more efficient and accurate. A simple precondition would transform the classic deconvolution-measured matching filter into a distribution, and since we are not modifying the modeled or measured data directly, the phase and amplitude of the original seismic signals are preserved and not altered. As we will show later, the comparison of the matching filter with the Dirac delta function maintains the convexity related to time shift.

In the following, we first review the matching-filter approach for FWI and then discuss the Radon-domain application to the matching-filter approach. After introducing the optimal-transport theory, we propose a new misfit function by comparing the resulting matching filter with the band-limited Dirac delta function based on the Wasserstein distance. In the example section, we first use a shifted wavelet to analyze the properties of the proposed
misfit function and compare it to other misfit functions. Then we use a modified Marmousi model and an offshore field data example to demonstrate the good performance of the proposed method in overcoming cycle skipping.

**FWI by matching filter**

The Radon transform, with its ability to highlight coherency in the data, can potentially benefit nonlinear misfit functions including the normalized global crosscorrelation, envelope, optimal-transport, and matching-filter approaches. In this study, however, we will focus on its features in combination with matching-filter-based FWI. Thus, we start with a brief review of the matching-filter approach.

Within the framework of FWI, we use an initial or available velocity model to simulate predicted data \( p \) to compare measured data \( d \). The fitting criteria form the optimization problem in which we minimize, for example, the L2-norm objective function:

\[
J = \frac{1}{2} \int \left[ (p(t) - d(t))^2 \right] dt.
\]  

The L2-norm misfit function measures the mismatch between two traces through a sample-by-sample comparison. As a result, it is prone to cycle skipping.

A strong remedy to such FWI limitations is to compute a matching filter that describes the difference between the measured and predicted data over the whole trace. Thus, a matching filter \( w \) satisfies:

\[
w(t) * d(t) = p(t),
\]

where \( \ast \) denotes the convolution operation. Equation 2 is a linear equation and can be solved effectively either in the time domain by deconvolution or in the frequency domain by division. When the velocity model is correct, the energy in the matching filter focuses to zero lag, more like an approximated Dirac delta function. Thus, we can formulate an optimization problem that penalizes the energy of the filter to zero lag. As our target for the resulting matching filter, its amplitude at near zero time lags, or, in other words, transport the energy of the filter to zero lag. As our target for the resulting matching filter is a band-limited Dirac delta function, we can use optimal-transport theory in designing such a misfit function.

\[
J = \frac{1}{2} \int \left[ P(t) w(t) \right] dt,
\]

where \( P(t) \) is the penalty function, e.g., \( P(t) = |t| \), will enhance energy beyond zero lag, and the objective is to minimize such enhancement.

**Matching filter in the Radon domain**

The matching filter computed using equation 2 takes the whole trace of the measured and predicted data as input and would suffer from crosstalk between different events. We use Radon transform to separate the events based on their space-time coherency, i.e., the slope information. This would lead to a more robust computation of the matching filter. For a common-shot gather, we apply Radon transform to both the modeled and observed data and then compute a matching filter in tau-\( p \) domain:

\[
\tilde{p} = \mathcal{R}[p], \tilde{d} = \mathcal{R}[d], \tilde{d} \ast \tilde{w} = \tilde{p}.
\]

Here, \( \mathcal{R} \) represents the forward Radon transform. In applications, we use local Radon transform (Wang et al., 2010) to capture the local coherence of seismic events. After transformation, the events should be separated based on their local slope. For example, in Figure 1, the red and blue correspond to different events in the measured and predicted data. If we compute the matching filter per trace by deconvolution of the measured and predicted data, the red and blue events would crosstalk with each other. Alternatively, the Radon-transform panels in Figures 1c and 1d show the blue and red events separated due to their distinct slope values. Deconvolution of the trace indexed by the same slope value would thus avoid the crosstalk in the original time-space domain and, as a result, lead to a more robust inversion process.

The other question is related to how we measure the focusing of the resulting matching filter. Equation 3 uses a penalty function. This penalty approach has a potential drawback: the misfit function is sensitive to the amplitude of the resulting matching filter. We can reduce the misfit value by decreasing the amplitude of the matching filter. However, the expected evolution of the matching filter should reduce its amplitude at large time lags while increasing its amplitude at near zero time lags, or, in other words, transport the energy of the filter to zero lag. As our target for the resulting matching filter is a band-limited Dirac delta function, we can use optimal-transport theory in designing such a misfit function.
naturally by measuring the Wasserstein distance between the resulting matching filter and the Dirac delta function.

**Wasserstein distance between matching filter and Dirac delta function**

Conventional optimal-transport approaches measure the Wasserstein distance between the predicted data \( p(t) \) and measured data \( d(t) \) directly. The difficulty in transforming the seismic data to a probability distribution limited its application (Metivier et al., 2018). We use the \( W^2 \) norm to measure the distance between the matching filter \( w(t) \) and the Dirac delta function (see Figure 2 for a demonstration of these two approaches). To fulfill the requirements of the optimal-transport theory, we need to transform \( w(t) \) to a distribution. We elect here to square and normalize it as follows:

\[
\omega'(t) = \frac{w^2(t)}{\int w^2(t) dt} = \frac{w^2}{\|w\|_2^2},
\]

Other kinds of preconditioning, such as the envelope, can be used. The modification of the matching filter to a probability distribution is important. First, it is a strict requirement for the application of the optimal-transport theory. Second, it makes the misfit function insensitive to the amplitude of the matching filter, and thus, avoids the problem faced by the conventional matching-filter misfit function of equation 3.

When the model parameters are accurate, the resulting matching filter reduces to a Dirac delta function or a band-limited version of it. This implies that the Dirac delta function \( \delta(t) \) is the target; thus, the misfit function can be formulated as

\[
J = W^2_2(\omega', \delta).
\]  

For the \( W^2 \) distance, readers can refer to Yang et al. (2018).

For the complete mathematical development of this proposed approach, such as the adjoint source computation, readers can see Sun and Alkhalifah (2018a).

**Misfit function convexity analysis**

In this section, we analyze the convexity properties of the discussed misfit functions. We will compare the least-squares \( L^2 \)-norm approach of equation 1, the conventional matching-filter approach of equation 3, conventional optimal transport with preconditioning using the affine scaling method (Yang and Engquist, 2018), and the proposed optimal-transport matching-filter approach of equation 6. For the proposed method, we use a Gaussian with a standard deviation of 0.008 s to represent the Dirac delta function.

In the first setup, as shown in Figure 3a, we set the measured data \( d(t) \) to a Ricker wavelet with a peak frequency of 10 Hz, while the predicted data \( p(t) \) is given by a time-shifted version of the measured data, i.e., \( p(t) = d(t + \tau) \). Here, \( \tau \) is the time shift ranging from -0.8 to 0.8 s. We show the comparison results for the different misfit functions in Figure 3b. We can observe that the \( L^2 \)-norm misfit function (red curve) and the optimal-transport approach with an affine scaling (green curve) have local minima. As discussed by Metivier et al. (2018), a precondition of the seismic signal by the affine scaling method artificially creates mass at each point. The optimal mass transport becomes more local instead of having global exchange properties between the initial and target signals. This makes the corresponding
misfit function less sensitive to the time shift and, therefore, destroys its convexity property. For the conventional matching-filter approach (the blue curve) and the proposed approach (the black curve), we obtain good convexity for this setup.

In an alternative setup, as shown in Figure 4a, we will not only shift the signal but also scale the amplitude of the signal. Now, the predicted data are given by the formula: \( p(t) = e^{-2\tau}d(t + \tau) \). As the time shift \( \tau \) ranges from -0.8 to 0.8 s, the scaling of the signal would vary from 0.2019 \( (e^{-0.8}) \) to 4.95 \( (e^{0.8}) \). Scaling plus shifting is reasonable as wave propagation in the subsurface experiences geometric spreading among other phenomena. Thus, waves arriving early would have relatively larger amplitudes than later arrivals. The performance of the misfit functions is shown in Figure 4b. Now we can see that the conventional matching-filter approach (the blue curve) loses its convexity due to its sensitivity to the amplitude, while the proposed method (the black curve) maintains its convexity property. In Figure 4c, we show a zoom of Figure 4b and focus on the part where the residual is small.

Through this simple but often-used test to measure the convexity of the objective function, we observe that compared to the conventional optimal-transport method with precondition by the affine scaling method, our proposed method shows better convexity. Compared with the conventional matching-filter approach, the proposed misfit function can ensure convexity, even if we scale and shift the signal.

Examples

The Marmousi model. The first example is the modified Marmousi model. As shown in Figure 5a, the true velocity model extends 2 km in depth and 8 km laterally. The initial velocity is a linearly increasing velocity with depth as shown in Figure 5b. The data set is modeled using 80 shots with a source interval of 100 m and 400 receivers with an interval of 20 m. The source is a Ricker wavelet with a 10 Hz peak frequency. We apply an absorbing boundary condition at the surface. In the inversion, we mute data below 3 Hz to verify that our proposed method is capable of overcoming the cycle-skipping problem without low frequencies. We apply time-domain inversions and utilize a low-pass filter allowing frequencies up to 10 Hz in the inversion. We do not sequentially increase the frequency as often implemented. Instead, we use the frequency range from 3 to 10 Hz simultaneously.
in the inversion. We use a Gaussian function with a constant standard deviation of 0.004 s to approximate the Dirac delta function in optimal-transport matching-filter misfit function of equation 6. We run the inversion over 200 iterations using a nonlinear conjugate gradient method.

The inverted result for the L2-norm misfit function is shown in Figure 6a. Due to cycle skipping, the result has very strong artifacts in the left part of the model, and it is generally far from the true model. Figure 6c is the result of the proposed Radon-domain optimal-transport matching-filter approach. Compared to the true model, the inverted model recovers both of the low-wavenumber and the fine-scale parts of the true model well, although there are some artifacts and reduced accuracy at the boundary due to the limited illumination. Compared to the conventional optimal-transport matching-filter result of Figure 6b, the Radon-domain approach obviously shows fewer artifacts and higher accuracy especially in the deeper part of the model. Considering the initial velocity is far away from the true one, the proposed method can mitigate cycle skipping and effectively converge toward the target model.

Figure 6. The inverted velocity using (a) the L2-norm misfit function, (b) the conventional optimal-transport matching-filter approach, and (c) the Radon-domain optimal-transport matching-filter approach. Due to cycle skipping, L2-norm misfit function failed to get a meaningful result. Compared to the result of (b) by the optimal-transport matching-filter approach in the time-space domain, the Radon-domain result of (c) shows higher accuracy, especially for the deeper part of the model.

The offshore field data set. The second example is a real marine data set from offshore Australia (Sun and Alkhalifah, 2018b). The offset range is from 160 to 8200 m. The initial velocity model converted from rms velocity is given in Figure 7a. We perform the inversion using the proposed misfit function with a low-pass filter applied to the data to cut frequencies below 3, 5, 10, 20, and 40 Hz, sequentially. During the inversion, total variation (TV) regularization (Alkhalifah et al., 2018; Esser et al., 2018) is used to reduce the noise. The inverted model is shown in Figure 7b. The updated model shows consistent structures and high resolution due to including high frequencies in the inversion as well as utilizing TV regularization. In the left panels of Figures 8a and 8b, we show one selected common-shot gather from the initial and inverted models. We compare it with the recorded shot gather at the same location in the right panel. Clearly, the inverted model reproduces the data that better matches the observed data, especially at the larger offsets where cycle skipping usually happens, and it is evident for the initial model. Considering the initial
velocity model is obtained from a crude rms velocity, we attribute the reasonably good result to the proposed misfit function’s ability to handle cycle-skipped data.

**Conclusion**

The measure of distance between the deconvolution-based matching filter and a Dirac delta function using optimal-transport theory provides a new framework for designing misfit functions for FWI. In combination with Radon-domain representation of observed and modeled data, the resulting form of the objective function is based on a solid mathematical foundation supported by optimal-transport theory. It also admits good results in the examples we shared. The current comparison between measured and modeled data using optimal-transport theory, though it may provide good results in some examples, as demonstrated is unnatural as data are not distributions and do not lend themselves easily to the requirements of optimal-transport theory. On the other hand, the matching filter, provided by a deconvolution of the measured data with the modeled data, is an ideal candidate for a distribution. The most valuable information embedded in the matching filter is the way it is distributed over time away from zero lag. The polarity information is less significant for our objective of transforming the filter to a Dirac delta function. As a result, we can easily transform the matching filter to a distribution, and this transformation maintains the convexity related to time shifts promised by the optimal-transport theory.

**Data and materials availability**

Data associated with this research are confidential and cannot be released.

Corresponding author: bingbing.sun@kaust.edu.sa

**References**


Sun, B., and T. Alkhalifah, 2018b, Automatic wave-equation migration velocity analysis by focusing subsurface virtual sources:


