Journal of Computational Physics 373 (2018) 188–209

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

The quadratic Wasserstein metric for earthquake location

Jing Chen, Yifan Chen, Hao Wu *, Dinghui Yang

Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 28 October 2017
Received in revised form 25 June 2018
Accepted 25 June 2018
Available online 9 July 2018

Keywords:
Optimal transport
Wasserstein metric
Inverse theory
Waveform inversion
Earthquake location

ABSTRACT

In Engquist et al. (2016) [8], the Wasserstein metric was successfully introduced to the full waveform inversion. We apply this method to the earthquake location problem. For this problem, the seismic stations are far from each other. Thus, the trace by trace comparison (Yang et al. [47]) is a natural way to compare the earthquake signals.

Under this framework, we have derived a concise analytic expression of the Frèchet gradient of the Wasserstein metric, which leads to a simple and efficient implementation of the adjoint method. We square and normalize the earthquake signals for comparison so that the convexity of the misfit function with respect to earthquake hypocenter and origin time can be realized and observed numerically. To reduce the impact of noise, which does not offset each other after the signals are squared, a new control parameter is introduced. Finally, the LMF (Levenberg–Marquardt–Fletcher) method is applied to solve the resulted optimization problem. According to the numerical experiments, only a few iterations are required to converge to the real earthquake hypocenter and origin time. Even for data with noise, we can obtain reasonable and convergent numerical results.

1. Introduction

The Wasserstein metric is an important concept in the optimal transport theory [33,40,41]. It measures the difference between two probability distributions by the optimal cost of rearranging one distribution into the other. This kind of problem was first proposed by French engineer Gaspard Monge [29], with the purpose of searching a way to move a pile of sand to a designated location at a minimum cost. As the metric provides a global comparison tool, it is very suitable to model and solve problems from computer vision [32], machine learning [2], etc.

From the mathematical point of view, there are many advantages of the Wasserstein metric [1,7,8,40], especially for the quadratic Wasserstein metric (W_2), e.g. the convexity with respect to shift, dilation and partial amplitude change, the insensitivity with respect to noise. In [7], Engquist and Froese first used this metric to measure the misfit between seismic signals. Then, the idea was developed to invert the velocity structure [8], in which due to the convexity of the Wasserstein metric, the full waveform inversion converges to the correct solution even from poor initial values. In [47], the method was further applied to more realistic examples. Motivated by this idea, Métivier and collaborators proposed the KR norm based full waveform inversion [26,27]. They also show the superiority of their method through several realistic problems. Different from Engquist's method, the KR norm is related to the Wasserstein metric with the linear cost function.

* Corresponding author.
E-mail addresses: jing-che16@mails.tsinghua.edu.cn (J. Chen), chenyifan14@mails.tsinghua.edu.cn (Y. Chen), hwu@tsinghua.edu.cn (H. Wu), dhyang@math.tsinghua.edu.cn (D. Yang).

https://doi.org/10.1016/j.jcp.2018.06.066
0021-9991/

In this study, we would like to apply the quadratic Wasserstein metric to the earthquake location problem, which is a fundamental challenge in geoscience [11,23,36] and finds plenty of applications in quantitative seismology [19,30,34,35]. This problem consists of two parts: the determination of hypocenter and origin time. In the early stage, the graphical method and simple grid searching method are used [28]. Thanks to the development of computational power and numerical methods, the problem can be solved by more effective iterative methods. In the past decades, the ray theory based earthquake location methods have been widely developed and applied in practice [11–13,38,42] because of its high efficiency and robustness. But this method is of low accuracy when the seismic wavelength is not small enough compared to the scale of wave propagation region [9,15,31,46]. To address this issue, the waveform based earthquake location methods [17,22,37,44,45] are developed to determine the earthquakes' parameters accurately. However, it is well known that the waveform inversion with ℓ^2 norm is suffering from the cycle-skipping problem, so it requires very accurate initial data for inversion. For the earthquake location problem, the situation is even worse, since the seismic focus is modeled by the highly singular delta function $\delta(\boldsymbol{x}-\boldsymbol{\xi})$ [3,23]. Fortunately, the Wasserstein metric has been shown to be effective in overcoming the famous cycle-skipping problem [8,26,27,47]. This is the motivation for us to work on this topic.

In the context of the earthquake location problems, the receivers are located far apart from each other. This is different from the situation that the receivers are close to each other in the problems of exploration seismology. Thus, we would like to follow the idea of the trace by trace comparison with W_2 metric [47]. In this paper, we apply this metric to invert the seismic focus parameters. A concise analytic expression of the Frèchet gradient is derived to simplify the numerical computation. We use the LMF method [10,20,24,25] to solve the optimization problem, which takes the advantages of the least square structure well. The numerical experiments show that the computational efficiency is greatly improved. We also would like to point out that the method developed in this paper may be effective for geological scale problems.

This paper is organized as follows. After reviewing the formulation and basic properties of the W_2 metric in Section 2, we apply it to the earthquake location problem in Section 3. Since we are considering the trace by trace comparison of the W_2 metric, it is very easy to derive the Fréchet gradient and the sensitivity kernel. The efficient LMF method is introduced to solve the optimization problem. In Section 4, the numerical experiments are provided to demonstrate the effectiveness and efficiency of the method. Finally, we conclude the paper in Section 5.

2. The quadratic Wasserstein metric

Let $\tilde{f}$ and $\tilde{g}$ be two probability density functions on $\mathbb{R}$, then the mathematical definition of the quadratic Wasserstein metric between $\tilde{f}$ and $\tilde{g}$ will be [7,8]:

$$W_2^2(\tilde{f},\tilde{g}) = \inf_{T\in\mathcal{M}} \int_{\mathbb{R}} |t-T(t)|^2\, \tilde{f}(t)\mathrm{d}t, \tag{2.1}$$

in which $\mathcal{M}$ is the set of all the rearrange maps from $\tilde{f}$ to $\tilde{g}$. According to the "Optimal transportation theorem for a quadratic cost on $\mathbb{R}$" (see P74 in [40]), the optimal transportation cost and the optimal map are given by

$$W_2^2(\tilde{f},\tilde{g}) = \int_0^1 \left|F^{-1}(t)-G^{-1}(t)\right|^2 \mathrm{d}t, \quad T(t)=G^{-1}(F(t)), \tag{2.2}$$

where $F(t)$ and $G(t)$ are the corresponding cumulative distribution functions of $\tilde{f}(t)$ and $\tilde{g}(t)$:

$$F(t)=\int_{-\infty}^{t}\tilde{f}(\tau)\mathrm{d}\tau, \quad G(t)=\int_{-\infty}^{t}\tilde{g}(\tau)\mathrm{d}\tau.$$

It is easy to verify that

$$\tilde{f}(t)=\tilde{g}(T(t))T'(t). \tag{2.3}$$

In this study, we prefer to use the 1D quadratic Wasserstein metric. This is based on the perspective that the signal at individual receivers can be treated as one-dimensional time series since these receivers are far from each other in geological scale problems. For the case of multiple receivers, we need to superimpose the contributions of each receiver, see Section 3 for details. In addition, the computation of the 1D quadratic Wasserstein metric can be easily implemented by formula (2.2).

In seismology, there are difficulties in applying the Wasserstein metric. Firstly, the seismic signal $f(t)$ and $g(t)$ are not positive. As an example, many seismograms at source have the form of Ricker wavelet

$$R(t)=A(1-2\pi^2 f_0^2 t^2)e^{-\pi^2 f_0^2 t^2}.$$

Here f_0 is the dominant frequency and A is the normalization factor. This Ricker wavelet is not always positive over the entire time axis. Secondly, the comparison between $\tilde{f}$ and $\tilde{g}$ under the Wasserstein metric requires the mass conservation, i.e.

$$\int_{\mathbb{R}} \tilde{f}(t)\mathrm{d}t = \int_{\mathbb{R}} \tilde{g}(t)\mathrm{d}t,$$

which is also not guaranteed for general seismic signals. Nevertheless, the above mentioned difficulties can be easily addressed by considering the following reformulated distance

$$\mathbf{d}(f,g) = W_2^2\left(\frac{f^2}{\langle f^2\rangle}, \frac{g^2}{\langle g^2\rangle}\right), \tag{2.4}$$

in which, the operator $\langle\cdot\rangle$ denote the integral over the real axis

$$\langle f\rangle = \int_{\mathbb{R}} f(t)\mathrm{d}t.$$

Remark 1. In Section 3.1 of manuscript [47], the authors prefer to add a constant c to ensure the positivity. However, in our numerical tests, the square strategy seems to be more suitable for the earthquake location problems.

Remark 2. In [8], the convexity of the quadratic Wasserstein metric with respect to shift, stretching and partial amplitude loss has been proved. Thus, we will not repeat here.

The aforementioned discussion is concerning the theoretical model but in practice, the signals $f(t)$ and $g(t)$ for consideration are often supported in $[0, t_f]$ for t_f large enough. Thus, the operator $\langle\cdot\rangle$ can be redefined as

$$\langle f(t)\rangle = \int_0^{t_f} f(t)\mathrm{d}t.$$

In the latter part of the paper, we will default to this notation unless otherwise specified.

2.1. The Fréchet gradient

We have defined the distance $\mathbf{d}(f,g)$ based on the Wasserstein metric. To solve the resulted optimization problem, it is necessary to derive the Fréchet gradient $\nabla_f \mathbf{d}$ first. Define

$$\mathcal{S}(f) = \frac{f^2}{\langle f^2\rangle}, \quad \mathcal{W}(\mathcal{F},\mathcal{G}) = W_2^2(\mathcal{F},\mathcal{G}),$$

then we have

$$\mathbf{d}(f,g) = \mathcal{W}(\mathcal{S}(f),\mathcal{S}(g)),$$

and

$$\nabla_f \mathbf{d} = \nabla_{\mathcal{F}}\mathcal{W}\cdot\nabla_f \mathcal{S}. \tag{2.5}$$

Before derivation, we have to emphasize that all the high order terms are ignored without any explanation. We first derive the gradient $\nabla_{\mathcal{F}}\mathcal{W}$. Let $\delta\mathcal{F}$ be a small perturbation[1] of $\mathcal{F}$, according to (2.1)–(2.2)

$$\mathcal{W}+\delta\mathcal{W} = \int_0^{t_f} |t-(T+\delta T)|^2 (\mathcal{F}+\delta\mathcal{F})\mathrm{d}t.$$

This leads to

$$\delta\mathcal{W} = \int_0^{t_f} |t-T|^2 \delta\mathcal{F}\mathrm{d}t - 2\int_0^{t_f} (t-T)\mathcal{F}\delta T\mathrm{d}t. \tag{2.6}$$

[1] In order to avoid repeating the explanation, we use δf to denote the small perturbation of arbitrary function f in the later part of the paper.

Since the Wasserstein metric measures the difference between two probability density functions, we can naturally assume that

$$\int_0^{t_f} \delta\mathcal{F}(t)\mathrm{d}t = 0, \tag{2.7}$$

Using the equation (2.3), we get

$$\mathcal{F} + \delta\mathcal{F} = \left(\mathcal{G}(T) + \mathcal{G}'(T)\delta T\right)(T' + (\delta T)'),$$

which yields

$$\delta\mathcal{F} = \mathcal{G}(T)(\delta T)' + \mathcal{G}'(T)T'\delta T = (\mathcal{G}(T)\delta T)'.$$

Integrating the above equation over $[0,t] \subset [0,t_f]$ leads to

$$\mathcal{G}(T(t))\delta T(t) = \int_0^t \delta\mathcal{F}(\tau)\mathrm{d}\tau + \mathcal{G}(T(0))\delta T(0) = \int_0^t \delta\mathcal{F}(\tau)\mathrm{d}\tau, \tag{2.8}$$

where the second equality holds since $T(0) = 0$ and

$$\mathcal{G}(T(0)) = \mathcal{G}(0) = \mathcal{S}(g(0)) = 0.$$

Using (2.8) in (2.6), we have, in light of (2.3):

$$\begin{aligned}\delta\mathcal{W} &= \int_0^{t_f} |t - T(t)|^2\, \delta\mathcal{F}\mathrm{d}t - 2\int_0^{t_f} (t - T(t))T'(t) \int_0^t \delta\mathcal{F}(\tau)\mathrm{d}\tau\mathrm{d}t \\ &= \int_0^{t_f} |t - T(t)|^2\, \delta\mathcal{F}\mathrm{d}t - 2\int_0^{t_f} \left(\int_t^{t_f} (\tau - T(\tau))T'(\tau)\mathrm{d}\tau\right) \delta\mathcal{F}(t)\mathrm{d}t = \int_0^{t_f} \varphi(t)\delta\mathcal{F}(t)\mathrm{d}t,\end{aligned} \tag{2.9}$$

where

$$\varphi(t) = |t - T(t)|^2 - 2\int_t^{t_f} (\tau - T(\tau))T'(\tau)\mathrm{d}\tau + C,$$

in which $C \in \mathbb{R}$ is a constant. Differentiating both sides of the above equation with respect to t gives

$$\varphi'(t) = 2(t - T(t)).$$

Thus, the simplest form of the function $\varphi(t)$ can be written as

$$\varphi(t) = 2\int_0^t (\tau - T(\tau))\mathrm{d}\tau. \tag{2.10}$$

Remark 3. We can also get this expression by the optimal transport theory and the duality theory of linear programming. For interested readers, we refer to P14 Theorem 1.17 in [33] for the details.

Next, we would like to derive $\nabla_f\mathcal{S}$. For small perturbation δf of f, we have

$$\mathcal{S} + \delta\mathcal{S} = \frac{(f+\delta f)^2}{\langle (f+\delta f)^2\rangle} = \frac{f^2 + 2f\delta f}{\langle f^2 + 2f\delta f\rangle} = \left(f^2 + 2f\delta f\right)\left(\frac{1}{\langle f^2\rangle} - \frac{2\langle f\delta f\rangle}{\langle f^2\rangle\langle f^2 + 2f\delta f\rangle}\right).$$

It follows that

$$\delta\mathcal{S} = \frac{2f\delta f}{\langle f^2\rangle} - \frac{2f^2\langle f\delta f\rangle}{\langle f^2\rangle^2}. \tag{2.11}$$

In summary, the Fréchet gradient can be obtained by combining (2.5) and (2.9)–(2.11)

$$\delta \mathbf{d} = \int_0^{t_f} \left(2 \int_0^t (\tau - T(\tau)) \mathrm{d}\tau \right) \left(\frac{2 f \delta f}{\langle f^2 \rangle} - \frac{2 f^2 \langle f \delta f \rangle}{\langle f^2 \rangle^2} \right) \mathrm{d}t = \int_0^{t_f} 4 \left(A(t) - B \right) f(t) \delta f(t) \mathrm{d}t, \tag{2.12}$$

where

$$A(t) = \frac{\int_0^t (\tau - T(\tau)) \mathrm{d}\tau}{\int_0^{t_f} f^2(t) \mathrm{d}t}, \quad B = \frac{\int_0^{t_f} \left(\int_0^t (\tau - T(\tau)) \mathrm{d}\tau \right) f^2(t) \mathrm{d}t}{\left(\int_0^{t_f} f^2(t) \mathrm{d}t \right)^2}.$$

2.2. Reduce the impact of noise

We now turn to discuss the impact of noise on the reformulated distance $\mathbf{d}(f, g)$. In (2.4), we take the square of the signals f, g to ensure the positivity, which implies that Theorem 3.1 in [8] is not directly applicable here. In the following, we consider a more general situation.

Theorem 1. *Let $\tilde{g}(t) : [0, 1] \to (0, M_1]$ and*

$$\tilde{f}_N(t) = \tilde{g}(t) + \tilde{r}_N(t),$$

$$\tilde{r}_N(t) = \begin{cases} \tilde{r}_1, & t \in [0, \frac{1}{N}], \\ \tilde{r}_2, & t \in (\frac{1}{N}, \frac{2}{N}], \\ \cdots & \\ \tilde{r}_N, & t \in (\frac{N-1}{N}, 1], \end{cases}$$

in which $\tilde{r}_j$ are i.i.d. random variables with zero expectation and bounded variance

$$\mathbb{E}\tilde{r}_j = 0, \quad \mathbb{D}\tilde{r}_j < +\infty, \quad j = 1, 2, \cdots, N.$$

We further assume that $\tilde{f}_N(t) : [0, 1] \to (0, M_2]$, then $\mathbb{E} W_2^2(\tilde{f}_N / \langle \tilde{f}_N \rangle, \tilde{g} / \langle \tilde{g} \rangle) = O(1/N)$.

The proofs are almost identical to Theorem 3.1 in [8]. Thus, they will not be reproduced here.

For practical problems, consider the signal $f_N(t)$, $g(t)$ defined on $[0, t_f]$ and

$$f_N(t) = g(t) + r_N(t). \tag{2.13}$$

Here

$$r_N(t) = \begin{cases} r_1, & t \in [0, \frac{t_f}{N}], \\ r_2, & t \in (\frac{t_f}{N}, \frac{2t_f}{N}], \\ \cdots & \\ r_N, & t \in (\frac{(N-1)t_f}{N}, t_f], \end{cases} \tag{2.14}$$

in which r_j are i.i.d. random variables with bounded expectation and variance

$$\mathbb{E}r_j = \mu, \quad \mathbb{D}r_j = \sigma^2, \quad j = 1, 2, \cdots, N. \tag{2.15}$$

Taking account of noise, we redefine the distance function in (2.4) as follows

$$\mathbf{d}_{\lambda(t)}(f_N, g) = W_2^2 \left(\frac{f_N^2}{\langle f_N^2 \rangle}, \frac{g^2 + \lambda}{\langle g^2 + \lambda \rangle} \right). \tag{2.16}$$

Here $\lambda = \lambda(t)$ is a given function of t satisfying

$$\lambda(t) + g^2(t) > 0, \quad t \in [0, t_f].$$

Let $\lambda(t) = 2\mu g(t) + \mu^2 + \sigma^2$, then we have

$$\begin{aligned} \mathbf{d}_{\lambda}(f_N, g) = W_2^2 \left(\frac{f_N^2}{\langle f_N^2 \rangle}, \frac{g^2 + \lambda}{\langle g^2 + \lambda \rangle} \right) &= W_2^2 \left(\frac{g^2 + 2 g r_N + r_N^2}{\langle g^2 + 2 g r_N + r_N^2 \rangle}, \frac{g^2 + \lambda}{\langle g^2 + \lambda \rangle} \right) \\ &= W_2^2 \left(\frac{(g^2 + \lambda) + (2 g r_N + r_N^2 - \lambda)}{\langle (g^2 + \lambda) + (2 g r_N + r_N^2 - \lambda) \rangle}, \frac{g^2 + \lambda}{\langle g^2 + \lambda \rangle} \right). \end{aligned}$$

Table 1
Example 2.1: the expectation values of the distance $\mathbb{E}\mathbf{d}_\lambda(f_N, g)$ with respect to λ and N. The last line is the expectation values of the L^2 distance between $f_N(t)$ and $g(t)$.

λ \ N	50	100	200	400	800
$0.8\lambda_*$	1.02×10^{-2}	8.43×10^{-3}	6.36×10^{-3}	5.56×10^{-3}	5.09×10^{-3}
$0.9\lambda_*$	8.65×10^{-3}	4.80×10^{-3}	2.79×10^{-3}	2.14×10^{-3}	1.64×10^{-3}
λ_*	7.42×10^{-3}	4.10×10^{-3}	2.09×10^{-3}	9.90×10^{-4}	5.34×10^{-4}
$1.1\lambda_*$	6.89×10^{-3}	4.78×10^{-3}	3.03×10^{-3}	1.99×10^{-3}	1.63×10^{-3}
$1.2\lambda_*$	1.03×10^{-2}	7.01×10^{-3}	5.31×10^{-3}	4.79×10^{-3}	4.04×10^{-3}
$\mathbb{E}\|f_N(t) - g(t)\|_2$	1.70×10^{-2}	1.66×10^{-2}	1.66×10^{-2}	1.67×10^{-2}	1.67×10^{-2}

Applying Theorem 1, we obtain

$$\mathbb{E}\mathbf{d}_\lambda(f_N, g) = O(\frac{1}{N}).$$

Remark 4. In many practical problems, $\mathbb{E}r_j = \mu = 0$. In such a case, $\lambda = \sigma^2$ is a constant independent of the variable t.

Example 2.1. In this example, we investigate the influence caused by the uniform distribution $r_j \sim U[-0.1, 0.1]$, $j = 1, 2, \cdots, N$. It follows that

$$\mathbb{E}r_j = 0, \quad \mathbb{D}r_j = \frac{1}{300}.$$

Let $g(t)$ be the Ricker wavelet $R(t - 2.5)$ with $A = 1$ and $f_0 = 2$ Hz. The signal $f_N(t)$ and the noise function $r_N(t)$ are given in (2.13)–(2.14). The time interval is $[0, 5]$. According to the above discussion, $\lambda_* = \frac{1}{300}$. In Table 1, we output the expectation values of the distance $\mathbb{E}\mathbf{d}_\lambda(f_N, g)$ with respect to the parameter λ and the number of time divisions N. For each configuration, we repeat 100 trials to compute the expectation values. For reference, we also output the expectation values of the L^2 distance between $f_N(t)$ and $g(t)$,

$$\|f_N(t) - g(t)\|_2 = \left(\int_0^5 |f_N(t) - g(t)|^2 \, \mathrm{d}t \right)^{1/2}.$$

From the table, we can see that $\mathbb{E}\mathbf{d}_\lambda(f_N, g) \approx O(\frac{1}{N})$ when $\lambda = \lambda_*$. This agrees with our theoretical discussion. Moreover, $\mathbb{E}\mathbf{d}_\lambda(f_N, g)$ decreases as N increases when λ is close to λ_*. On the other hand, the expectation values of the L^2 distance remains unchanged.

Example 2.2. In this example, we investigate the influence caused by the normal distribution $r_j \sim N[0, 0.1^2]$, $j = 1, 2, \cdots, N$. It follows that

$$\mathbb{E}r_j = 0, \quad \mathbb{D}r_j = 0.01.$$

Let $g(t)$ be the Ricker wavelet $R(t - 2.5)$ with $A = 1$ and $f_0 = 2$ Hz. The signal $f_N(t)$ and the noise function $r_N(t)$ are given in (2.13)–(2.14). The time interval is $[0, 5]$. According to the above discussion, $\lambda_* = \frac{1}{100}$. In Table 2, we output the expectation values of the distance $\mathbb{E}\mathbf{d}_\lambda(f_N, g)$ with respect to the parameter λ and the number of time divisions N. For each configuration, we repeat 100 trials to compute the expectation values. For reference, we also output the expectation values of the L^2 distance between $f_N(t)$ and $g(t)$. From the table, we can draw the same conclusion as in Example 2.1.

The aforementioned discussions and experiments point out that the parameter λ should be specified in order to reduce the impact of noise. This requires us to estimate the mean and variance of the noise, which will cost some extra efforts. Fortunately, there are many statistical methods to achieve this goal, e.g. Independent Component Analysis [14]. Moreover, the numerical experiments imply that the estimation does not need to be sufficiently accurate.

3. The application to earthquake location

Up to now, we have proposed the reformulated distance (2.4) to measure two earthquake signals and studied its properties. Next, we would like to apply this distance to determine the real earthquake hypocenter $\boldsymbol{\xi}_T$ and the origin time τ_T. Its mathematical formulation is written as follows

$$(\boldsymbol{\xi}_T, \tau_T) = \underset{\boldsymbol{\xi}, \tau}{\operatorname{argmin}} \sum_r \chi_r(\boldsymbol{\xi}, \tau), \tag{3.1}$$

Table 2
Example 2.2: the expectation values of the distance $\mathbb{E}\mathbf{d}_\lambda(f_N, g)$ with respect to λ and N. The last line is the expectation values of the L^2 distance between $f_N(t)$ and $g(t)$.

λ \ N	50	100	200	400	800
$0.8\lambda_*$	4.77×10^{-2}	2.73×10^{-2}	1.69×10^{-2}	1.53×10^{-2}	1.06×10^{-2}
$0.9\lambda_*$	4.45×10^{-2}	2.30×10^{-2}	1.40×10^{-2}	7.28×10^{-3}	4.97×10^{-3}
λ_*	3.74×10^{-2}	2.01×10^{-2}	1.26×10^{-2}	6.30×10^{-3}	3.00×10^{-3}
$1.1\lambda_*$	3.42×10^{-2}	2.40×10^{-2}	1.33×10^{-2}	8.54×10^{-3}	4.28×10^{-3}
$1.2\lambda_*$	3.86×10^{-2}	2.52×10^{-2}	1.56×10^{-2}	9.73×10^{-3}	9.11×10^{-3}
$\mathbb{E}\,\|f_N(t)-g(t)\|_2$	4.89×10^{-2}	4.99×10^{-2}	5.07×10^{-2}	5.04×10^{-2}	4.99×10^{-2}

where the misfit function at the r-th receiver $\chi_r(\boldsymbol{\xi},\tau)$ is defined by

$$\chi_r(\boldsymbol{\xi},\tau)=\mathbf{d}(d_r(t),s(\boldsymbol{\eta}_r,t)). \tag{3.2}$$

In equation (3.1), the misfit functions are summed up so that the information of all the receivers are taken into account. The real earthquake signal $d_r(t)$ and the synthetic earthquake signal $s(\boldsymbol{x},t)$ can be considered as the solution

$$d_r(t)=u(\boldsymbol{\eta}_r,t;\boldsymbol{\xi}_T,\tau_T),\quad s(\boldsymbol{x},t)=u(\boldsymbol{x},t;\boldsymbol{\xi},\tau), \tag{3.3}$$

of the acoustic wave equation initial-boundary value problem

$$\frac{\partial^2 u(\boldsymbol{x},t;\boldsymbol{\xi},\tau)}{\partial t^2}=\nabla\cdot\Big(c^2(\boldsymbol{x})\nabla u(\boldsymbol{x},t;\boldsymbol{\xi},\tau)\Big)+R(t-\tau)\delta(\boldsymbol{x}-\boldsymbol{\xi}),\quad \boldsymbol{x},\boldsymbol{\xi}\in\Omega, \tag{3.4}$$

$$u(\boldsymbol{x},0;\boldsymbol{\xi},\tau)=\partial_t u(\boldsymbol{x},0;\boldsymbol{\xi},\tau)=0,\quad \boldsymbol{x}\in\Omega, \tag{3.5}$$

$$\boldsymbol{n}\cdot\Big(c^2(\boldsymbol{x})\nabla u(\boldsymbol{x},t;\boldsymbol{\xi},\tau)\Big)=0,\quad \boldsymbol{x}\in\partial\Omega. \tag{3.6}$$

In the above equations, $c(\boldsymbol{x})$ denotes the wave speed and $\boldsymbol{\eta}_r$ is the location of the r-th receiver. The computational domain Ω is a subset of the d-dimensional real Euclidean space $\mathbb{R}^d$ and $\boldsymbol{n}$ is the outward unit normal vector to the domain Ω. In this study, the seismic rupture is modeled by the point source $\delta(\boldsymbol{x}-\boldsymbol{\xi})$ since its scale is much smaller compared to the scale of seismic wave propagated [3,23]. We also remark that the reflection boundary condition (3.6) is used to simplify the model. There is no essential difficulty to consider other boundary conditions, e.g. the perfectly matched layer absorbing boundary condition [18].

Remark 5. In practice, the real signal is superimposed with noise

$$d_r(t)=u(\boldsymbol{\eta}_r,t;\boldsymbol{\xi}_T,\tau_T)+r_N(t).$$

Thus, we prefer to use the distance given in (2.16) to define the misfit function

$$\chi_r(\boldsymbol{\xi},\tau)=\mathbf{d}_\lambda(d_r(t),s(\boldsymbol{\eta}_r,t)). \tag{3.7}$$

3.1. The adjoint method

The perturbation of earthquake hypocenter $\delta\boldsymbol{\xi}\ll 1$ and origin time $\delta\tau\ll 1$ would generate the perturbation of wave function

$$\delta s(\boldsymbol{x},t)=u(\boldsymbol{x},t;\boldsymbol{\xi}+\delta\boldsymbol{\xi},\tau+\delta\tau)-u(\boldsymbol{x},t;\boldsymbol{\xi},\tau). \tag{3.8}$$

According to (3.4)–(3.6), $\delta s(\boldsymbol{x},t)$ satisfies the acoustic wave equation

$$\frac{\partial^2\delta s(\boldsymbol{x},t)}{\partial t^2}=\nabla\cdot\Big(c^2(\boldsymbol{x})\nabla\delta s(\boldsymbol{x},t)\Big)+R(t-(\tau+\delta\tau))\delta(\boldsymbol{x}-(\boldsymbol{\xi}+\delta\boldsymbol{\xi}))-R(t-\tau)\delta(\boldsymbol{x}-\boldsymbol{\xi}),\quad \boldsymbol{x},\boldsymbol{\xi}\in\Omega, \tag{3.9}$$

$$\delta s(\boldsymbol{x},0)=\partial_t\delta s(\boldsymbol{x},0)=0,\quad \boldsymbol{x}\in\Omega, \tag{3.10}$$

$$\boldsymbol{n}\cdot\Big(c^2(\boldsymbol{x})\nabla\delta s(\boldsymbol{x},0)\Big)=0,\quad \boldsymbol{x}\in\partial\Omega. \tag{3.11}$$

Multiply an arbitrary test function $w_r(\boldsymbol{x},t)$ on equation (3.9), integrate it on $\Omega\times[0,t_f]$ and use integration by parts to obtain

$$\begin{aligned}
&\int_0^{t_f}\int_\Omega \frac{\partial^2 w_r}{\partial t^2}\delta s \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \frac{\partial w_r}{\partial t}\delta s\Big|_{t=t_f}\mathrm{d}\boldsymbol{x} + \int_\Omega w_r\frac{\partial \delta s}{\partial t}\Big|_{t=t_f}\mathrm{d}\boldsymbol{x}\\
&= \int_0^{t_f}\int_\Omega \delta s\nabla\cdot(c^2\nabla w_r)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_0^{t_f}\int_{\partial\Omega}\boldsymbol{n}\cdot(c^2\nabla w_r)\delta s\mathrm{d}\zeta\mathrm{d}t + \int_0^{t_f} R(t-(\tau+\delta\tau))w_r(\boldsymbol{\xi}+\delta\boldsymbol{\xi},t) - R(t-\tau)w_r(\boldsymbol{\xi},t)\mathrm{d}t\\
&\approx \int_0^{t_f}\int_\Omega \delta s\nabla\cdot(c^2\nabla w_r)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_0^{t_f}\int_{\partial\Omega}\boldsymbol{n}\cdot(c^2\nabla w_r)\delta s\mathrm{d}\zeta\mathrm{d}t + \int_0^{t_f} R(t-\tau)\nabla w_r(\boldsymbol{\xi},t)\cdot\delta\boldsymbol{\xi} - R'(t-\tau)w_r(\boldsymbol{\xi},t)\delta\tau\mathrm{d}t.
\end{aligned} \tag{3.12}$$

In the last step, the Taylor expansion is used and higher order terms are ignored.

On the other hand, the misfit function (3.2) also generates the perturbation with respect to $\delta s(\boldsymbol{x},t)$. Assuming $\|\delta s(\boldsymbol{x},t)\| \ll 1$ and taking into account of (2.12), we have

$$\begin{aligned}
\delta\chi_r &= \chi_r(\boldsymbol{\xi}+\delta\boldsymbol{\xi},\tau+\delta\tau) - \chi_r(\boldsymbol{\xi},\tau)\\
&\approx \int_0^{t_f} 4\,(A(t)-B)\,s(\boldsymbol{\eta}_r,t)\delta s(\boldsymbol{\eta}_r,t)\mathrm{d}t\\
&= \int_0^{t_f}\int_\Omega 4\,(A(t)-B)\,s(\boldsymbol{\eta}_r,t)\delta s(\boldsymbol{x},t)\delta(\boldsymbol{x}-\boldsymbol{\eta}_r)\mathrm{d}\boldsymbol{x}\mathrm{d}t,
\end{aligned} \tag{3.13}$$

where "$\approx$" is obtained by ignoring high order terms of $\delta s(\boldsymbol{x},t)$.

Let $w_r(\boldsymbol{x},t)$ satisfy the adjoint equation

$$\frac{\partial^2 w_r(\boldsymbol{x},t)}{\partial t^2} = \nabla\cdot\Big(c^2(\boldsymbol{x})\nabla w_r(\boldsymbol{x},t)\Big) + 4\,(A(t)-B)\,s(\boldsymbol{\eta}_r,t)\delta(\boldsymbol{x}-\boldsymbol{\eta}_r),\quad \boldsymbol{x},\boldsymbol{\xi}\in\Omega, \tag{3.14}$$

$$w_r(\boldsymbol{x},t_f) = \frac{\partial w_r(\boldsymbol{x},t_f)}{\partial t} = 0,\quad \boldsymbol{x}\in\Omega, \tag{3.15}$$

$$\boldsymbol{n}\cdot\Big(c^2(\boldsymbol{x})\nabla w_r(\boldsymbol{x},t)\Big) = 0,\quad \boldsymbol{x}\in\partial\Omega. \tag{3.16}$$

Thus, the relation between $\delta\chi_r$ and $\delta\boldsymbol{\xi}$, $\delta\tau$ can be obtained by adding (3.12) to (3.13)

$$\delta\chi_r = K_r^{\boldsymbol{\xi}}\cdot\delta\boldsymbol{\xi} + K_r^{\tau}\delta\tau, \tag{3.17}$$

in which the sensitivity kernel for the hypocenter $\boldsymbol{\xi}$ and the origin time τ is

$$K_r^{\boldsymbol{\xi}} = \int_0^{t_f} R(t-\tau)\nabla w_r(\boldsymbol{\xi},t)\mathrm{d}t, \tag{3.18}$$

$$K_r^{\tau} = -\int_0^{t_f} R'(t-\tau)w_r(\boldsymbol{\xi},t)\mathrm{d}t. \tag{3.19}$$

3.2. The LMF method

According to (2.2), (2.4) and (3.1)–(3.11), the earthquake location is formulated as a least square optimization problem. Therefore, we can consider some special methods to improve the convergence. Through a large number of numerical tests, we found that the LMF method [10,20,24,25] works very well. In the following, we briefly review the basic idea of the algorithm.

In order to be consistent with the literature of optimization theory, all the symbols and notations in this subsection are independent from the other part of the paper. The general form of the least-square problem can be written as

$$\min f(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{m} r_i^2(\boldsymbol{x}),\quad \boldsymbol{x}\in\mathbb{R}^n,\ m\geq n, \tag{3.20}$$

where the residual function

$$\vec{r}(\boldsymbol{x}) = (r_1(\boldsymbol{x}), r_2(\boldsymbol{x}), \cdots, r_m(\boldsymbol{x}))^T \in \mathbb{R}^m.$$

The gradient of f is

$$\nabla f(\boldsymbol{x}) = \sum_{i=1}^{m} r_i(\boldsymbol{x}) \nabla r_i(\boldsymbol{x}) = J(\boldsymbol{x})^T \vec{r}(\boldsymbol{x}), \tag{3.21}$$

in which $J(\boldsymbol{x})$ is the Jacobi matrix of $\vec{r}(\boldsymbol{x})$

$$J(\boldsymbol{x}) = (\nabla r_1, \nabla r_2, \cdots, \nabla r_m)^T \in \mathbb{R}^{m \times n}.$$

The key ingredient of LMF method is

$$\left(J_k^T J_k + \nu_k I\right) \boldsymbol{d}_k = -J_k^T \vec{r}_k, \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k, \tag{3.22}$$

where I is the identity matrix and $\nu_k \geq 0$ is a parameter in each iteration step. It is introduced to improve the convergence and efficiency. To adjust this parameter, we define

$$\gamma_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{d}_k)}{q_k(0) - q_k(\boldsymbol{d}_k)}, \tag{3.23}$$

with

$$q_k(\boldsymbol{d}) = \frac{1}{2} \left(J_k \boldsymbol{d} + \vec{r}_k\right)^T \left(J_k \boldsymbol{d} + \vec{r}_k\right). \tag{3.24}$$

Now the detailed implementation of the LMF method is summarized below.

Algorithm 1 *(The LMF method).*

1. Set tolerance value $\varepsilon = 0.01$, the break-off step $K = 20$ and $\mu = 2$. Let $k = 0$ and give the initial value $\boldsymbol{x}_0$. Thus, we have the initial adjustable parameter $\nu_0 = 10^{-6} \times \max\left|\text{diag}(J_0^T J_0)\right|$.
2. For $\boldsymbol{x}_k$ and ν_k, solve the equation (3.22) to obtain $\boldsymbol{d}_k$. And we can calculate γ_k using equation (3.23).
3. If $\gamma_k > 0$, let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$, $\nu_{k+1} = \nu_k \times \max\{\frac{1}{3}, 1 - (2\gamma_k - 1)^3\}$, $\mu = 2$ and $k = k + 1$. If $\|f(\boldsymbol{x}_k)\| < \varepsilon$, output $\boldsymbol{x}_k$ and stop.
4. If $\gamma_k < 0$, let $\nu_k = \mu \nu_k$, $\mu = 2\mu$.
5. If $k > K$, output the error message: "The iteration doesn't converge." and stop. Otherwise, go to step 2 for another iteration.

In the above algorithm, we require that the objective function always decreases. Otherwise, we will decrease the radius of the trust region. For the noise-free situation, the idea works well since the objective function has nice convexity in a large area. However, due to the influence of data noise, the optimization objective function will appear some local minimum. Therefore, it may be not suitable to require the objective function decreasing during the whole iteration. Thus, we modify the LMF method as follows.

Algorithm 2 *(The modified LMF method for noise data).*

1. Set tolerance value $\varepsilon = 0.01$, the break-off step $K = 20$ and $\mu = 2$. Let $k = 0$ and give the initial value $\boldsymbol{x}_0$. Thus, we have the adjustable parameter $\nu_0 = 10^{-3}$ and its upper limit $\eta = 10^{-3}$.
2. For $\boldsymbol{x}_k$ and ν_k, solve the equation (3.22) to obtain $\boldsymbol{d}_k$. And we can calculate γ_k using equation (3.23).
3. If $\gamma_k > 0$ or $\nu_k \geq \eta$, let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$, $\nu_{k+1} = \min\{\eta, \nu_k \times \max\{\frac{1}{3}, 1 - (2\gamma_k - 1)^3\}\}$, $\mu = 2$ and $k = k + 1$. If $\|f(\boldsymbol{x}_k)\| < \varepsilon$, output $\boldsymbol{x}_k$ and stop.
4. If $\gamma_k < 0$ and $\nu_k < \eta$, let $\nu_k = \min\{\eta, \mu \nu_k\}$, $\mu = 2\mu$.
5. If $k > K$, output the error message: "The iteration doesn't converge." and stop. Otherwise, go to step 2 for another iteration.

Remark 6. The criteria $\|f(\boldsymbol{x}_k)\| < \varepsilon$ in the above algorithm may not be reached because of the data noise. For this problem, we have two options. One is to change ε, but this could be a little tricky to choose an appropriate parameter ε. The other is to wait until the iteration ends. Then, we can choose the numerical solutions corresponding to the smallest value of the objective function

$$k_* = \underset{k}{\text{argmin}} \|f(\boldsymbol{x}_k)\|.$$

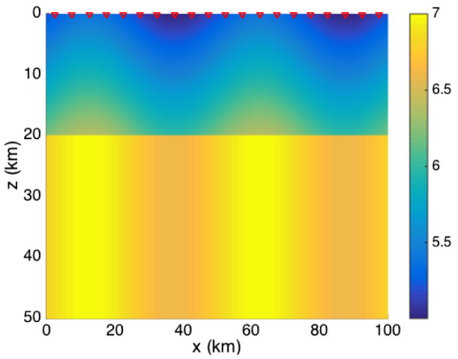

Fig. 1. Illustration of two-layer model. The red triangles indicate the receivers. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

4. Numerical experiments

In this section, two examples are presented to demonstrate the validity of the optimal transport model. We first study the convexity of the optimization objective function with respect to the initial earthquake hypocenter. The convergence and efficiency of Algorithm 1 will be discussed for the situation when the seismic signals are noise-free. For the situation of noisy data, the convergence of Algorithm 2 are numerically investigated.

To solve the acoustic wave equation (3.4), we simply use the finite difference schemes [6,21,47]. Inside the earth, the perfectly matched layer boundary condition [18] is applied to absorb the outgoing waves. On the other hand, the reflection boundary condition (3.6) is used to model the free surface of the earth. To discretize the delta function for the point source $\delta(\boldsymbol{x}-\boldsymbol{\xi})$, we borrow the idea from [43]. It writes

$$\delta(x)=\begin{cases}\frac{1}{h}\left(1-\frac{5}{4}\left|\frac{x}{h}\right|^2-\frac{35}{12}\left|\frac{x}{h}\right|^3+\frac{21}{4}\left|\frac{x}{h}\right|^4-\frac{25}{12}\left|\frac{x}{h}\right|^5\right), & |x|\le h,\\ \frac{1}{h}\left(-4+\frac{75}{4}\left|\frac{x}{h}\right|-\frac{245}{8}\left|\frac{x}{h}\right|^2+\frac{545}{24}\left|\frac{x}{h}\right|^3-\frac{63}{8}\left|\frac{x}{h}\right|^4+\frac{25}{24}\left|\frac{x}{h}\right|^5\right), & h<|x|\le 2h,\\ \frac{1}{h}\left(18-\frac{153}{4}\left|\frac{x}{h}\right|+\frac{255}{8}\left|\frac{x}{h}\right|^2-\frac{313}{24}\left|\frac{x}{h}\right|^3+\frac{21}{8}\left|\frac{x}{h}\right|^4-\frac{5}{24}\left|\frac{x}{h}\right|^5\right), & 2h<|x|\le 3h,\\ 0, & |x|>3h.\end{cases}$$

Here h is a numerical parameter which is related to the mesh size.

4.1. The two-layer model

Consider the two-layer model in the bounded domain $\Omega=[0,100\text{ km}]\times[0,50\text{ km}]$, the wave speed is

$$c(x,z)=\begin{cases}5.2+0.05z+0.2\sin\frac{\pi x}{25}, & 0\text{ km}\le z\le 20\text{ km},\\ 6.8+0.2\sin\frac{\pi x}{25}, & z>20\text{ km}.\end{cases}$$

The unit is 'km/s'. The computational time interval $I=[0,35\text{ s}]$. The dominant frequency of the earthquakes is $f_0=2$ Hz. There are 20 equidistant receivers on the surface

$$\boldsymbol{\eta}_r=(x_r,z_r)=(5r-2.5\text{ km},0),\quad r=1,2,\cdots,20,$$

see Fig. 1 for illustration.

First, we output the cross-section of the optimization objective function

$$\Psi(\boldsymbol{\xi})=\sum_r \chi_r(\boldsymbol{\xi},\tau_T),$$

in which $\chi_r(\boldsymbol{\xi},\tau)$ is defined in (3.2). Here, the distance between the real signal $d_r(t)$ and the synthetic signal $s(\boldsymbol{\eta}_r,t)$ is measured by the quadratic Wasserstein metric of the normalized square signals (QWN_2)

$$W_2^2\left(\frac{d_r^2(t)}{\langle d_r^2(t)\rangle},\frac{s^2(\boldsymbol{\eta}_r,t)}{\langle s^2(\boldsymbol{\eta}_r,t)\rangle}\right).$$

As a comparison, we also output the corresponding objective function under other types of distance as follows:

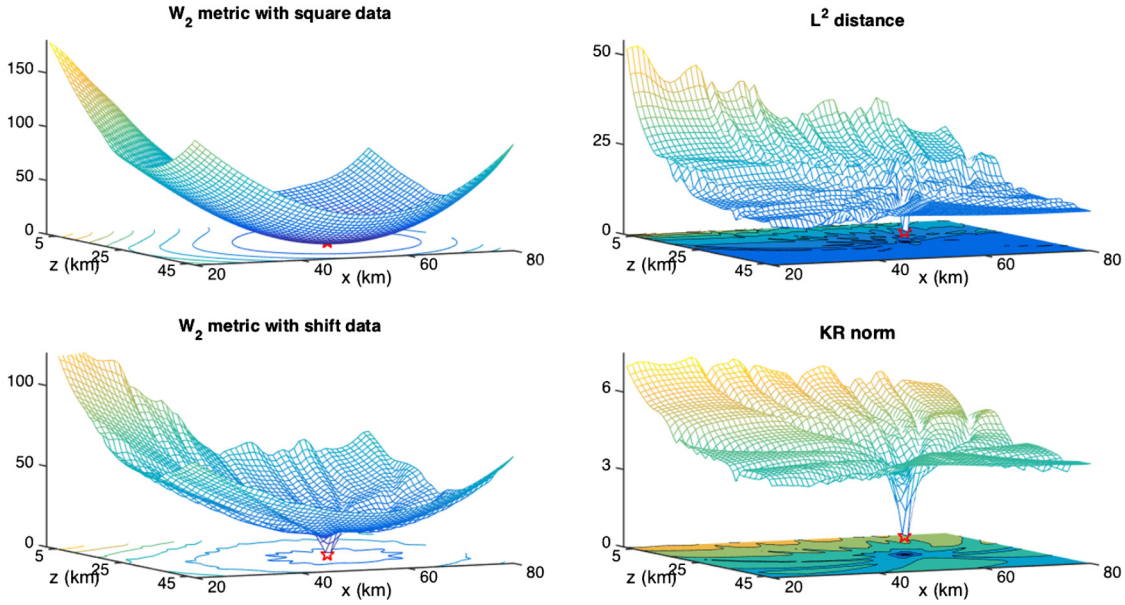

Fig. 2. The two-layer model, the cross-section of the optimization objective function with respect to different measures, case (i). The red pentagram denotes the real earthquake hypocenter.

- The relative L^2 distance (RLD):

$$\frac{\int_0^{t_f} \left|d_r(t) - s(\boldsymbol{\eta}_r, t)\right|^2 \mathrm{d}t}{\int_0^{t_f} |d_r(t)|^2 \,\mathrm{d}t};$$

- The quadratic Wasserstein metric of the normalized shift signals (QWN_c) [47]:

$$W_2^2\left(\frac{d_r(t)+c}{\langle d_r(t)+c\rangle}, \frac{s(\boldsymbol{\eta}_r,t)+c}{\langle s(\boldsymbol{\eta}_r,t)+c\rangle}\right),$$

  here c is a constant to ensure the positive;
- The Kantorovich–Rubinstein norm of the original signals (KRN) [16,26]:

$$W_1(d_r(t), s(\boldsymbol{\eta}_r,t)) = \max_{\varphi\in \mathrm{BLip}_1} \int_0^{t_f} \varphi(t)\left(d_r(t) - s(\boldsymbol{\eta}_r,t)\right)\mathrm{d}t,$$

  in which BLip_1 is the space of bounded 1-Lipschitz functions, such that

$$(i)\ \forall (t_1,t_2)\in[0,t_f],\ |\varphi(t_1)-\varphi(t_2)| \le |t_1-t_2|, \quad (ii)\ \forall t\in[0,t_f],\ |\varphi(t)|\le 1.$$

Here we consider two different earthquake hypocenters, one is below the Moho discontinuity (Fig. 2) and another is above the Moho discontinuity (Fig. 3):

(i) $\boldsymbol{\xi}_T = (57.604\ \mathrm{km}, 26.726\ \mathrm{km}), \quad \tau_T = 10.184\ \mathrm{s}.$

(ii) $\boldsymbol{\xi}_T = (46.234\ \mathrm{km}, 7.124\ \mathrm{km}), \quad \tau_T = 10.782\ \mathrm{s};$

From these figures, we can observe nice convexity property of the optimization objective function $\Psi(\boldsymbol{\xi})$ with respect to earthquake hypocenter $\boldsymbol{\xi} = (\zeta_x, \zeta_z)$ by QWN_2. For other distances, it seems the convexity property is not good enough.

Next, we test the LMF method (Algorithm 1) using 200 experiments. The real and initial earthquake hypocenter $\boldsymbol{\xi}_T^i$, $\boldsymbol{\xi}^i$ are both uniformly distributed over [20 km, 80 km] × [3 km, 40 km], and the real and initial original time τ_T^i, τ^i are both uniformly distributed over [7.5 s, 12.5 s]. Their spatial distributions and the histograms of the distance between the real and the initial hypocenters

$$d^i = \left\|\boldsymbol{\xi}_T^i - \boldsymbol{\xi}^i\right\|_2,$$

are presented in Fig. 4. For all the methods, we randomly select seven receivers for inversion, e.g. $r = 4, 5, 7, 9, 12, 14, 18$. In Table 3, we can see the convergence results for the LMF method, the Gauss–Newton (GN) method, and the BFGS method.

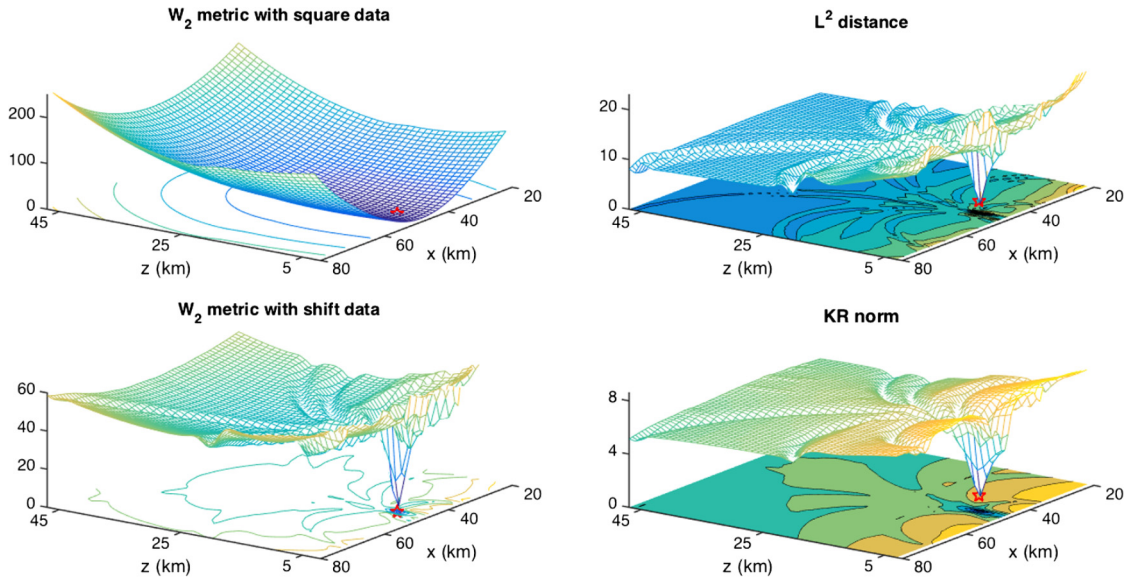

Fig. 3. The two-layer model, the cross-section of the optimization objective function with respect to different measures, case (ii). The red pentagram denotes the real earthquake hypocenter.

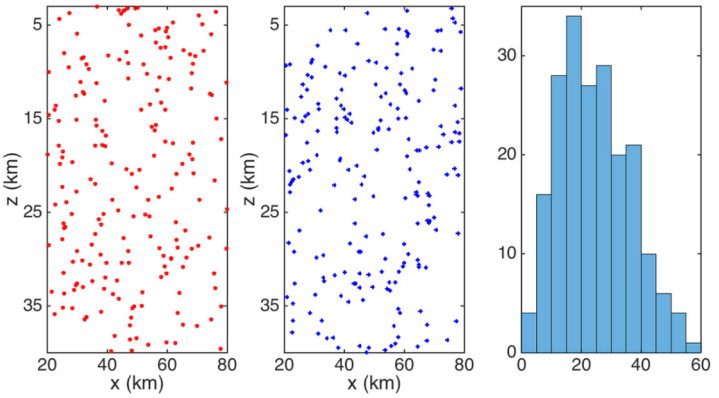

Fig. 4. The two-layer model. Left: the spatial distribution of the real earthquake hypocenter $\boldsymbol{\xi}_T^i$; Middle: the spatial distribution of the initial earthquake hypocenter $\boldsymbol{\xi}^i$; Right: the distance distribution histogram between the real and the initial earthquake hypocenter d^i.

Table 3
The two-layer model. Convergent results for the LMF method, the GN method and the BFGS method.

	Correct convergence	Divergence	Error convergence	Total
LMF	200	0	0	200
GN	147	53	0	200
BFGS	190	0	10	200

Precisely, it is shown that the LMF method correctly converges in all the tests, while there are 53 divergent results by the GN method and 10 error convergence results by the BFGS method. For the convergent cases, we output the mean and standard deviation of iterations for the three methods in Table 4. It is obvious to see that the LMF method converges faster than the BFGS method. Considering all the above factors, we can conclude that the LMF method is a better choice here.

Then, we output the convergent history of the LMF method (Algorithm 1) by two special examples. Their parameters are selected as follows:

$$(i)\ \boldsymbol{\xi}_T = (57.604\ \text{km},\ 26.726\ \text{km}),\ \tau_T = 10.184\ \text{s},\quad \boldsymbol{\xi} = (32.653\ \text{km},\ 12.214\ \text{km}),\ \tau = 12.108\ \text{s};$$

$$(ii)\ \boldsymbol{\xi}_T = (46.234\ \text{km},\ 13.124\ \text{km}),\ \tau_T = 10.782\ \text{s},\quad \boldsymbol{\xi} = (59.572\ \text{km},\ 29.013\ \text{km}),\ \tau = 9.908\ \text{s};$$

In Fig. 5, we can see the convergent trajectories, the absolute errors of the earthquake hypocenter and the value of Wasserstein distance. These figures show that the method converges to the real earthquake hypocenter very quickly.

Table 4
The two-layer model. Mean and Standard Deviation of iterations for the LMF method, the GN method and the BFGS method.

	Mean of iterations	Standard Deviation of iterations
LMF	5.93	1.90
GN	5.59	1.71
BFGS	10.80	2.84

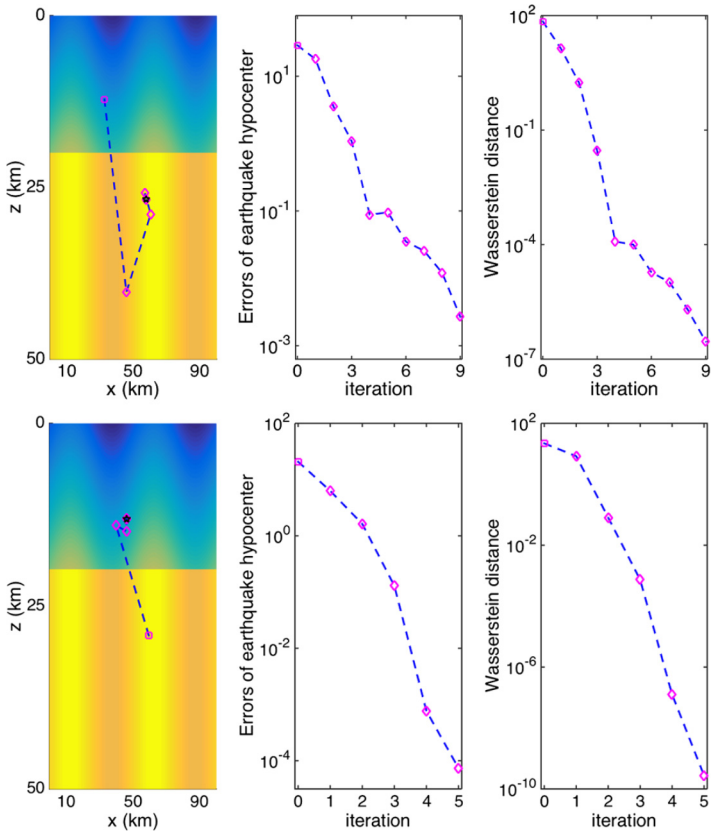

Fig. 5. Convergence history of the two-layer model. Up for case (i), and Down for case (ii). Left: the convergent trajectories; Mid: the absolute errors between the real and computed earthquake hypocenter with respect to iteration steps; Right: the Wasserstein distance between the real and synthetic earthquake signals with respect to iteration steps. The magenta square is the initial hypocenter, the magenta diamond denotes the hypocenter in the iterative process, and the black pentagram is the real hypocenter.

At last, we test the effectiveness of the new method for the noisy data. The same parameters (i) and (ii) are selected here. The real earthquake signal can be regarded as

$$d_r(t) = u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T) + N_r(t).$$

Here $N_r(t)$ is subject to the normal distribution with mean $\mu = 0$ and the standard deviation

$$\sigma = R \times \max_t \left| u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T) \right|.$$

The ratio R will be selected as 5%, 10%, 15% and 20% respectively in the later tests. These signals are illustrated in Fig. 6. Obviously, a time window that contains the main part of $u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T)$ can be chosen to reduce the impact of noise. As discussed above, we use the formulation (3.7) in Remark 5 and the modified LMF method (Algorithm 2) to deal with the noise. The convergent histories are output in Fig. 7–8. From these figures, we can see the location errors and the misfit functions oscillate during the iteration. And the iteration step k_* which corresponds to the smallest value of the misfit function does not correspond to the smallest location error. These are the unavoidable effects of noise. In Table 5–6, we output k_*, the corresponding misfit value and the corresponding location error. We can see the location results are still good enough. In addition, we notice that the locations error may be reduced as the noise ratio R increase. This is also caused by the randomness of the noise. Nevertheless, the above results show a strong adaptability to the noise of the new model and method.

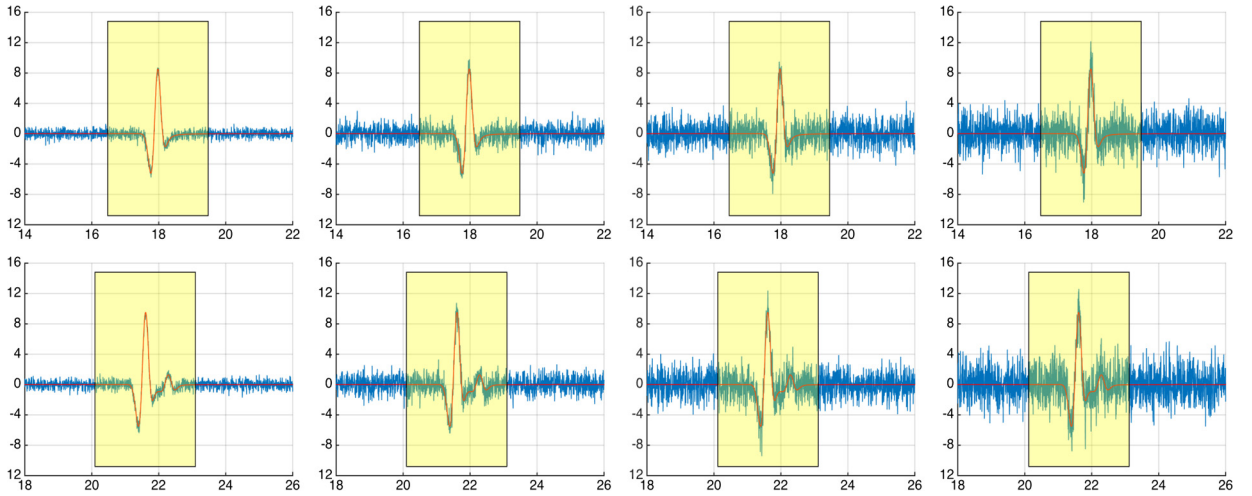

Fig. 6. Illustration of signal with noise in the two-layer model. The signal with noise $d_r(t)$ (blue line) and the noise free signal $u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T)$ for receiver $r = 7$. The horizontal axis is the time t. Up: parameters (i); Down: parameters (ii); From left to right, the ratio $R = 5\%$, 10%, 15%, 20% respectively.

Table 5
The two-layer model with noise data, case (i). The smallest misfit value, the corresponding iteration step k_* and the location error.

R	k_*	The misfit value	The location error (km)
5%	9	3.74×10^{-3}	7.80×10^{-2}
10%	3	1.37×10^{-2}	3.10×10^{-1}
15%	10	3.44×10^{-2}	4.08×10^{-1}
20%	12	4.45×10^{-2}	2.85×10^{-1}

Table 6
The two-layer model with noise data, case (ii). The smallest misfit value, the corresponding iteration step k_* and the location error.

R	k_*	The misfit value	The location error (km)
5%	15	4.33×10^{-3}	1.42×10^{-1}
10%	6	1.06×10^{-2}	2.02×10^{-1}
15%	3	1.76×10^{-2}	2.31×10^{-2}
20%	3	3.55×10^{-2}	2.22×10^{-1}

Table 7
The subduction plate model: the horizontal positions of receivers, with unit 'km'.

r	1	2	3	4	5	6	7	8	9	10	11	12
x_r	21	33	39	58	68	74	86	98	126	132	158	197

4.2. The subduction plate model

Let us consider a typical seismogenic zone model here [37,39]. It consists of the crust, the mantle, and the undulating Moho discontinuity. In addition, there is a subduction zone with a thin low-velocity layer atop a fast velocity layer in the mantle. The earthquake may occur in any of these areas. Taking into account the complex velocity structure, it is much difficult to locate the earthquake. In the simulating domain $\Omega = [0,\ 200\ \text{km}] \times [0,\ 200\ \text{km}]$, the wave speed is

$$c(x,z) = \begin{cases} 5.5, & 0 < z \le 33 + 5\sin\frac{\pi x}{40}, \\ 7.8, & 33 + 5\sin\frac{\pi x}{40} < z \le 45 + 0.4x, \\ 7.488, & 45 + 0.4x < z \le 60 + 0.4x, \\ 8.268, & 60 + 0.4x < z \le 85 + 0.4x, \\ 7.8, & \text{others}. \end{cases}$$

with unit 'km/s'. There are 12 randomly distributed receivers $\boldsymbol{\eta}_r = (x_r, z_r)$ on the surface $z_r = 0$ km. In Table 7, we output their horizontal positions. This velocity model is illustrated in Fig. 9. The dominant frequency of the earthquake is $f_0 = 2$ Hz and the simulating time interval $I = [0,\ 55\ \text{s}]$.

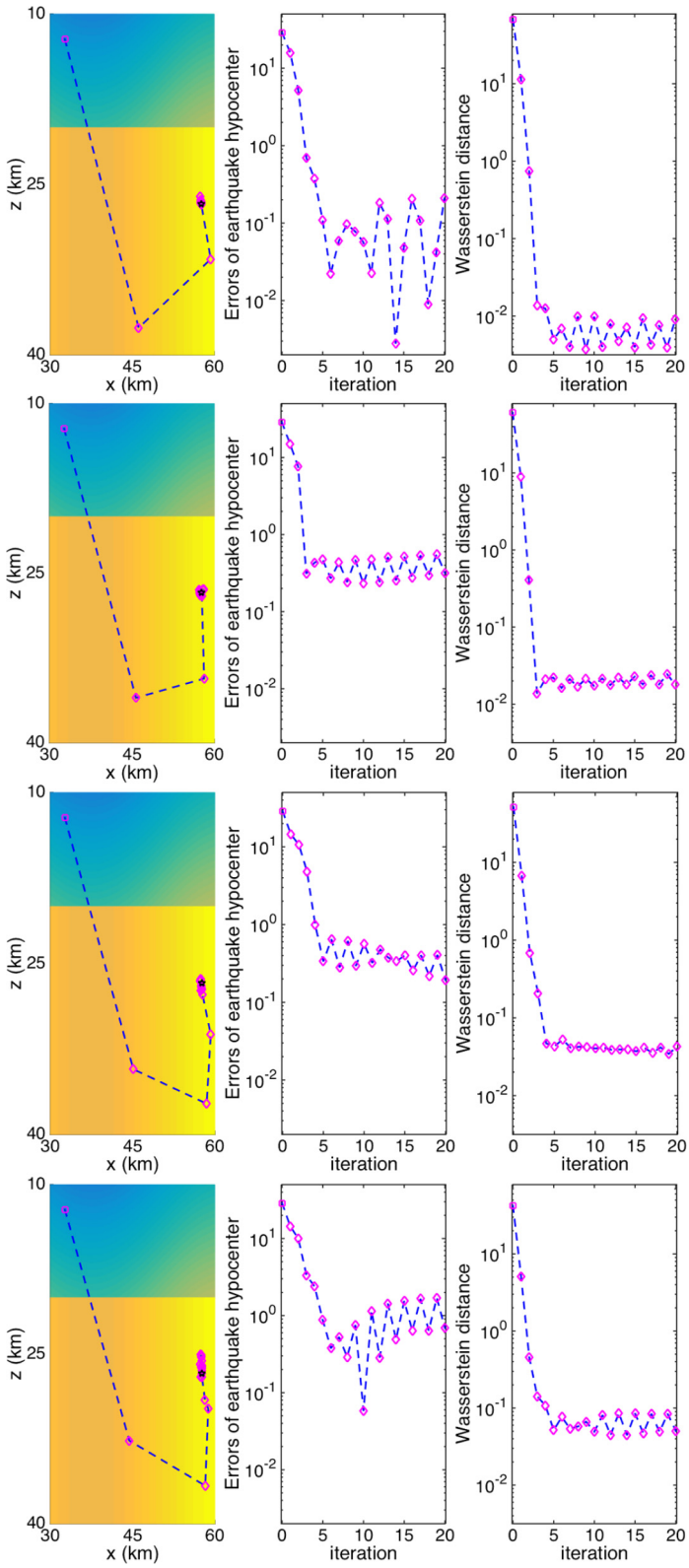

Fig. 7. Convergence history of the two-layer model with noise data, case (i). From up to bottom, the ratio $R = 5\%$, 10%, 15%, 20% respectively. Left: the convergent trajectories; Mid: the absolute errors between the real and computed earthquake hypocenter with respect to iteration steps; Right: the Wasserstein distance between the real and synthetic earthquake signals with respect to iteration steps. The magenta square is the initial hypocenter, the magenta diamond denotes the hypocenter in the iterative process, and the black pentagram is the real hypocenter.

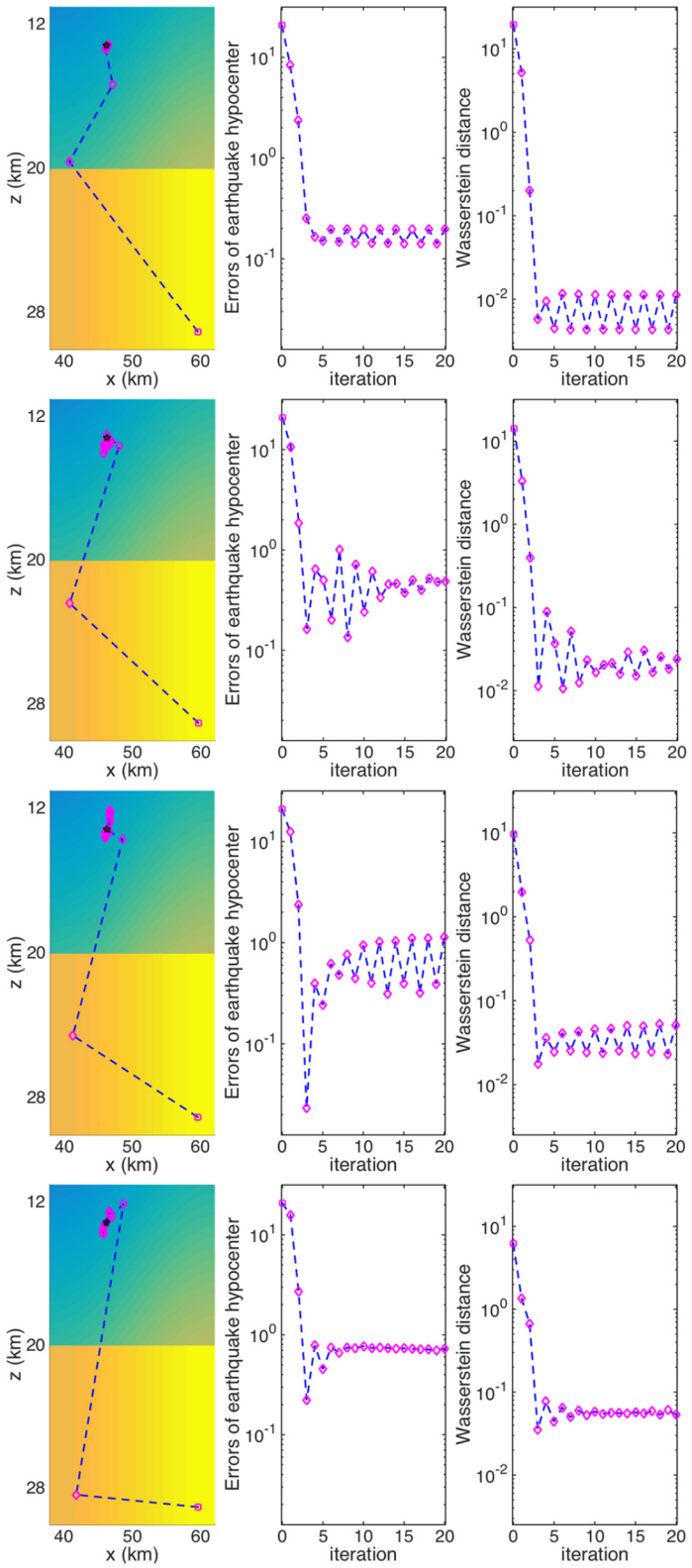

Fig. 8. Convergence history of the two-layer model with noise data, case (ii). From up to bottom, the ratio $R = 5\%$, 10%, 15%, 20% respectively. Left: the convergent trajectories; Mid: the absolute errors between the real and computed earthquake hypocenter with respect to iteration steps; Right: the Wasserstein distance between the real and synthetic earthquake signals with respect to iteration steps. The magenta square is the initial hypocenter, the magenta diamond denotes the hypocenter in the iterative process, and the black pentagram is the real hypocenter.

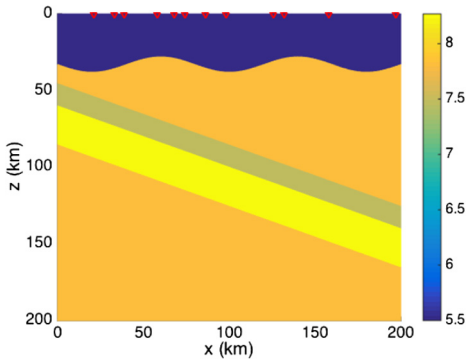

Fig. 9. Illustration of the subduction plate model. The read triangles indicate the receivers.

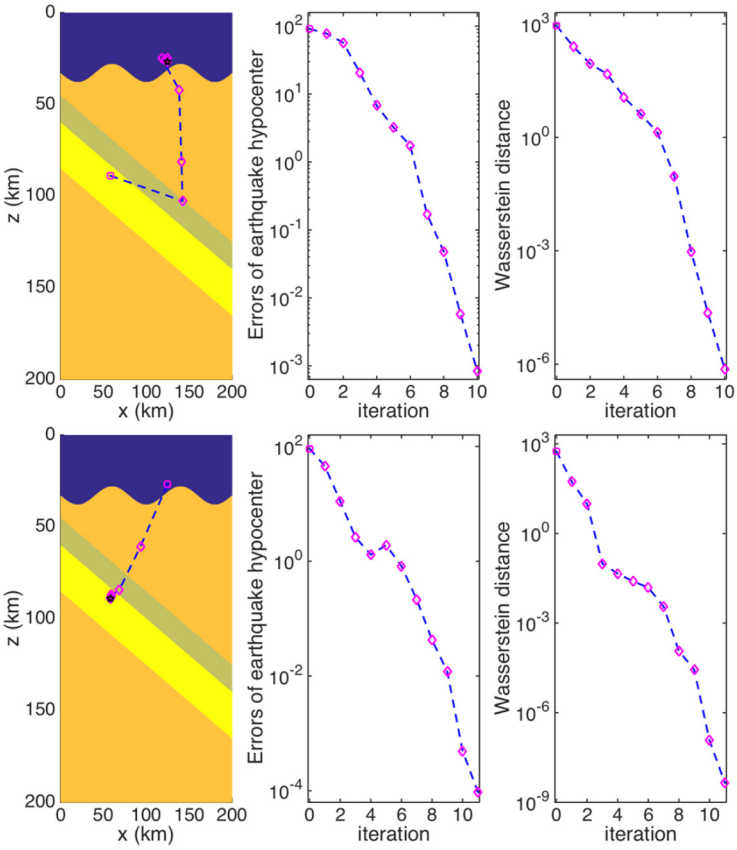

Fig. 10. Convergence history of the subduction plate model. Up for case (i), and Down for case (ii). Left: the convergent trajectories; Mid: the absolute errors between the real and computed earthquake hypocenter with respect to iteration steps; Right: the Wasserstein distance between the real and synthetic earthquake signals with respect to iteration steps. The magenta square is the initial hypocenter, the magenta diamond denotes the hypocenter in the iterative process, and the black pentagram is the real hypocenter.

First, consider the ideal situation that there is no data noise. We investigate the case when the earthquake occurs in the crust but the initial hypocenter of the earthquake is chosen in the subduction zone. Its contrary case is also taken into account. The parameters are selected as follows:

$$(i)\ \boldsymbol{\xi}_T = (124.694\ \text{km},\ 26.762\ \text{km}),\ \tau_T = 5.00\ \text{s},\quad \boldsymbol{\xi} = (58.056\ \text{km},\ 88.985\ \text{km}),\ \tau = 6.79\ \text{s};$$

$$(ii)\ \boldsymbol{\xi}_T = (58.056\ \text{km},\ 88.985\ \text{km}),\ \tau_T = 6.79\ \text{s},\quad \boldsymbol{\xi} = (124.694\ \text{km},\ 26.762\ \text{km}),\ \tau = 5.00\ \text{s}.$$

The convergent trajectories, absolute errors of the earthquake hypocenter and the value of Wasserstein distance are output in Fig. 10. From it, we can observe nice convergence property of the new method.

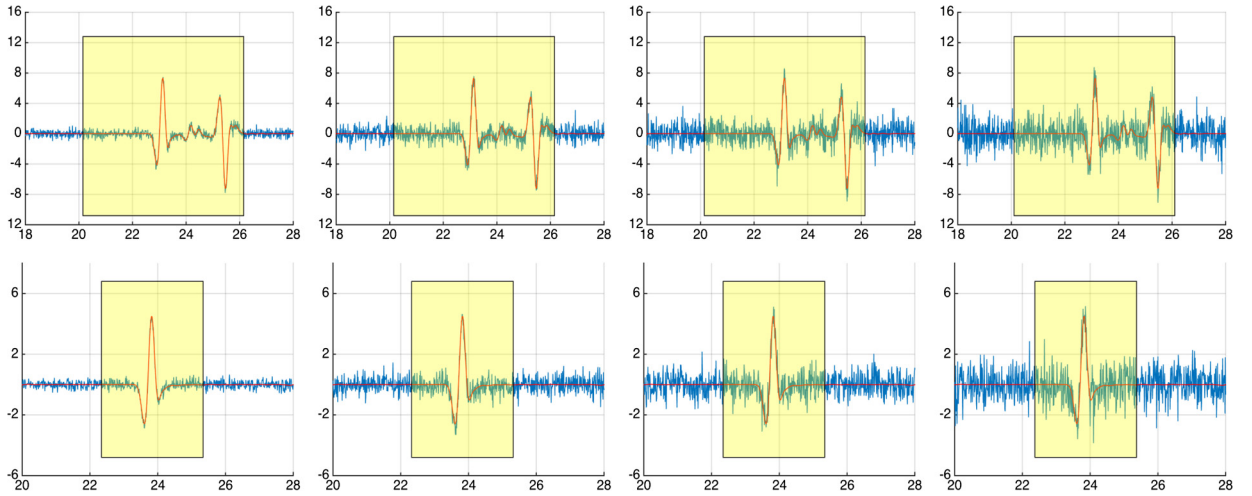

Fig. 11. Illustration of signal with noise in the subduction plate model. The signal with noise $d_r(t)$ (blue line) and the noise free signal $u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T)$ for receiver $r = 4$. The horizontal axis is the time t. Up: parameters (i); Down: parameters (ii); From left to right, the ratio $R = 5\%$, 10%, 15%, 20% respectively.

Table 8
The subduction plate mode with noise data, case (i). The smallest misfit value, the corresponding iteration step k_* and the location error.

R	k_*	The misfit value	The location error (km)
5%	4	4.49×10^{-1}	5.41×10^{-2}
10%	15	5.67×10^{-1}	3.96×10^{-1}
15%	6	4.15×10^{-1}	3.93×10^{-1}
20%	12	4.45×10^{-1}	1.24×10^{-1}

Table 9
The subduction plate mode with noise data, case (ii). The smallest misfit value, the corresponding iteration step k_* and the location error.

R	k_*	The misfit value	The location error (km)
5%	7	8.56×10^{-3}	2.40×10^{-1}
10%	19	2.70×10^{-2}	2.92×10^{-1}
15%	6	4.96×10^{-2}	2.10×10^{-1}
20%	18	7.64×10^{-2}	5.99×10^{-1}

We next consider the signal containing noise. We select the same parameters (i) and (ii). The noise is added to the real earthquake signals in the same way as in Subsection 4.1. In Fig. 11, the real earthquake signal with noise $d_r(t)$ and the noise-free signal $u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T)$ are presented. In order to reduce the impact of noise, it is necessary and reasonable to select a time window that contains the main part of $u(\boldsymbol{\eta}_r, t; \boldsymbol{\xi}_T, \tau_T)$. Moreover, the technique proposed in Remark 5 and the modified LMF method (Algorithm 2) are also applied here. In Fig. 12–13, we output the convergent history. As discussed in the previous Subsection, the location errors and the misfit functions oscillate due to the noise effect. Thus, it is impossible to get accurate location results. A reasonable choice is to select the iteration step k_* which corresponds to the smallest value of the misfit function. These values are output in Table 8–9. From it, we can see the location results are more satisfying by the new model and method.

5. Conclusion

In this paper, we apply the quadratic Wasserstein metric to the earthquake location problem. The numerical evidence suggests that the convexity of the misfit function with respect to the earthquake hypocenter and origin time, based on the quadratic Wasserstein metric, is much better than the one based on the L^2 metric. This makes it possible to accurately locate the earthquakes even starting from very far initial values. Besides, since the misfit function is close to the quadratic function, the LMF method could be a good choice to solve the resulted optimization problem. According to our numerical tests, the LMF method has obvious advantages over the GN method and the BFGS method. When the original signal is affected by noise, we make a little modification of the quadratic Wasserstein metric based misfit function and the LMF method. In all the numerical experiments, the location errors are smaller than 1 km. These location results, according to our best knowledge, are pretty good.

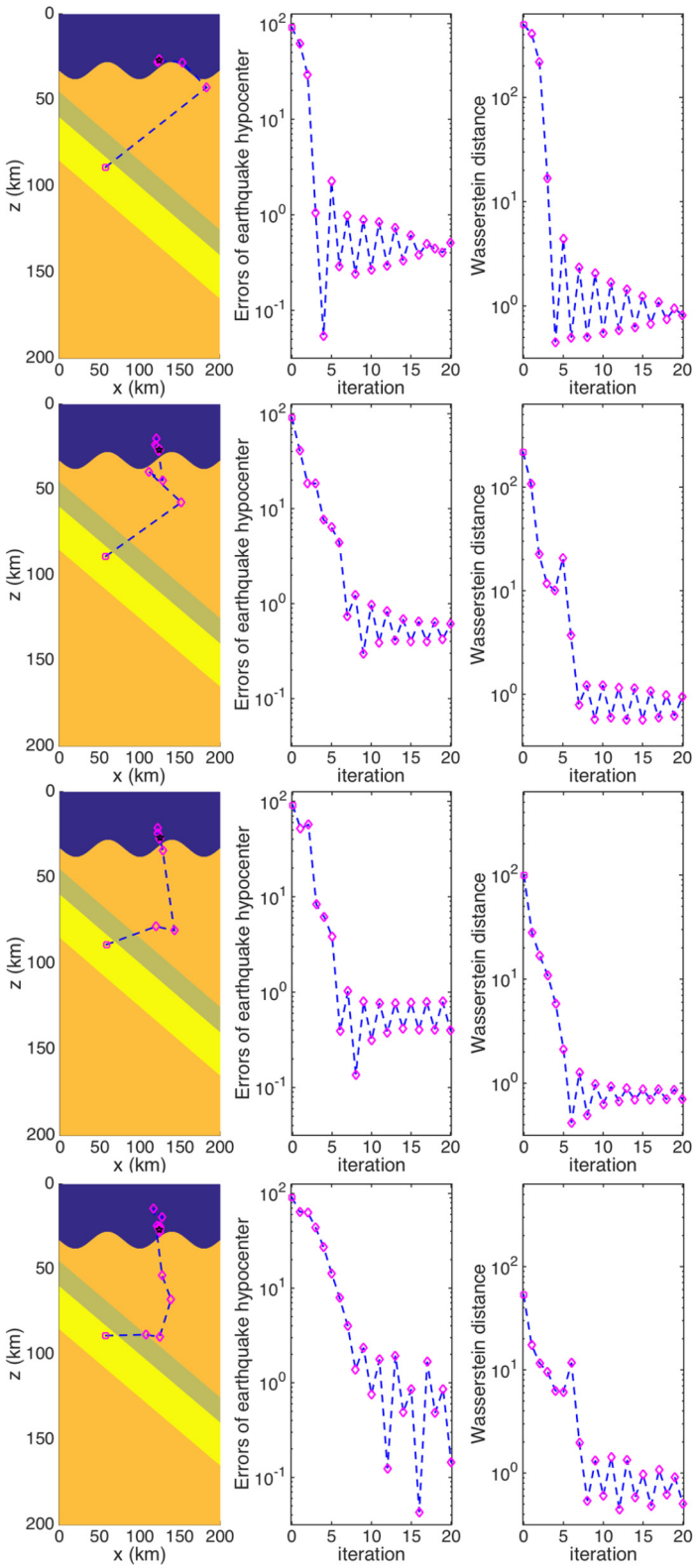

Fig. 12. Convergence history of the subduction plate model with noise data, case (i). From up to bottom, the ratio $R = 5\%$, 10%, 15%, 20% respectively. Left: the convergent trajectories; Mid: the absolute errors between the real and computed earthquake hypocenter with respect to iteration steps; Right: the Wasserstein distance between the real and synthetic earthquake signals with respect to iteration steps. The magenta square is the initial hypocenter, the magenta diamond denotes the hypocenter in the iterative process, and the black pentagram is the real hypocenter.

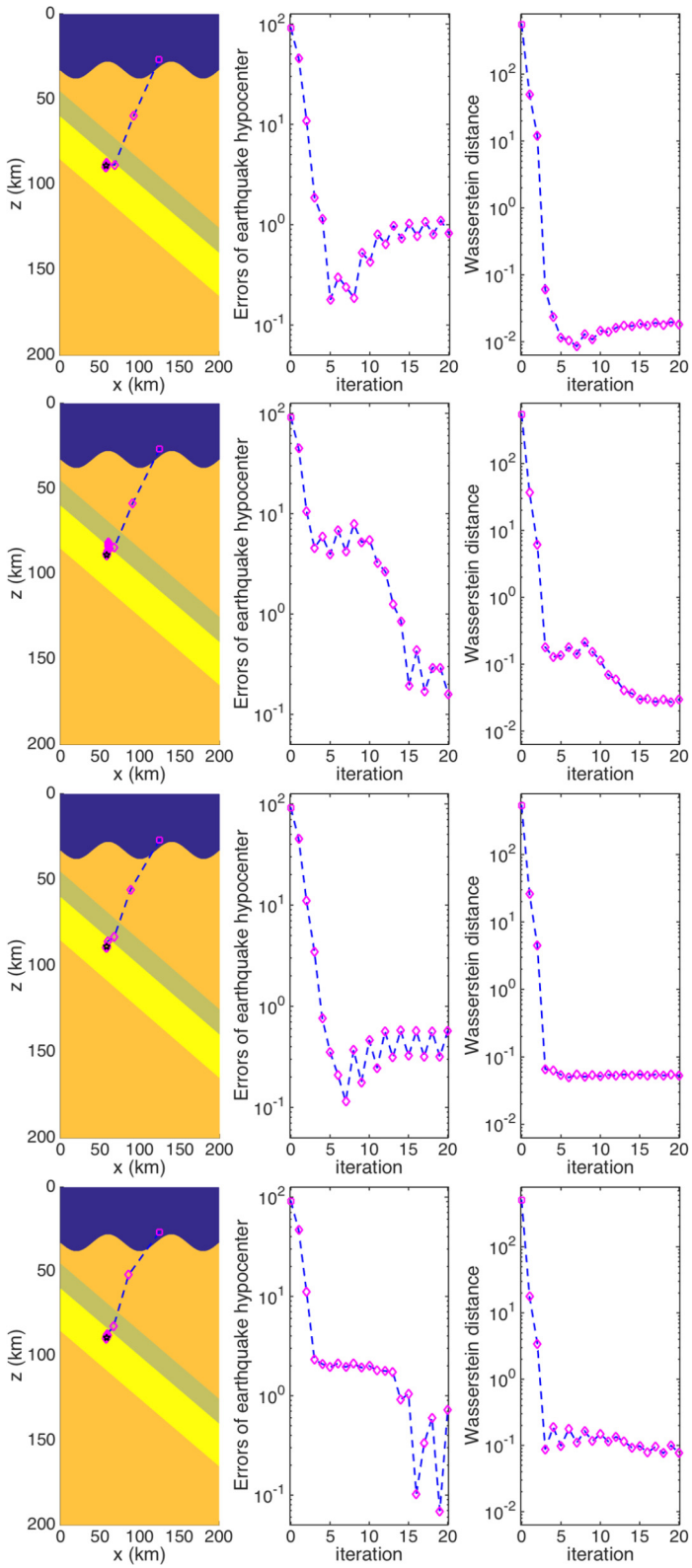

Fig. 13. Convergence history of the subduction plate model with noise data, case (ii). From up to bottom, the ratio $R = 5\%$, 10%, 15%, 20% respectively. Left: the convergent trajectories; Mid: the absolute errors between the real and computed earthquake hypocenter with respect to iteration steps; Right: the Wasserstein distance between the real and synthetic earthquake signals with respect to iteration steps. The magenta square is the initial hypocenter, the magenta diamond denotes the hypocenter in the iterative process, and the black pentagram is the real hypocenter.

We also need to point out that, since both the real and synthetic signals have been normalized, it is not possible to determine the amplitude of the seismogram at source. To deal with this difficulty, we may need to introduce the unbalanced optimal transport theory [4,5]. Moreover, the techniques developed in this paper may be applicable to the inversion of many micro-earthquakes [19,30]. We are currently working on these interesting topics and hope to report these in subsequent papers.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 41390452, 11101236, 91730306), the National Key R&D Program on Monitoring, Early Warning and Prevention of Major Natural Disaster (Grant No. 2017YFC1500301), and SRF for ROCS, SEM. The authors are grateful to Prof. Shi Jin for his inspiration and helpful suggestions and discussions that greatly improve the presentation. Hao Wu would like to acknowledge Prof. Björn Engquist and Prof. Brittany D. Froese for their valuable comments.

References

[1] L. Ambrosio, N. Gigli, A user guide to optimal transport, in: Modelling and Optimisation of Flows on Networks, Springer, 2013, pp. 1–155.
[2] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017, pp. 214–223.
[3] K. Aki, P.G. Richards, Quantitative Seismology: Theory and Methods, vol. II, W.H. Freeman & Co (Sd), 1980.
[4] L. Chizat, G. Peyré, B. Schmitzer, F.X. Vialard, An interpolating distance between optimal transport and Fisher–Rao metrics, Found. Comput. Math. (2016), https://doi.org/10.1007/s10208-016-9331-y.
[5] L. Chizat, G. Peyré, B. Schmitzer, F.X. Vialard, Scaling algorithms for unbalanced optimal transport problems, Math. Comput. (2018), https://doi.org/10.1090/mcom/3303.
[6] M.A. Dablain, The application of high-order differencing to the scalar wave equation, Geophysics 51 (1) (1986) 54–66.
[7] B. Engquist, B.D. Froese, Application of the Wasserstein metric to seismic signals, Commun. Math. Sci. 12 (5) (2014) 979–988.
[8] B. Engquist, B.D. Froese, Y.N. Yang, Optimal transport for seismic full waveform inversion, Commun. Math. Sci. 14 (8) (2016) 2309–2330.
[9] B. Engquist, O. Runborg, Computational high frequency wave propagation, Acta Numer. 12 (2003) 181–266.
[10] R. Fletcher, Practical Methods of Optimization, Second edition, John Wiley and Sons, Chichester, 1991.
[11] M.C. Ge, Analysis of source location algorithms, part I: overview and non-iterative methods, J. Acoust. Emiss. 21 (2003) 14–28.
[12] M.C. Ge, Analysis of source location algorithms, part II: iterative methods, J. Acoust. Emiss. 21 (2003) 29–51.
[13] L. Geiger, Probability method for the determination of earthquake epicenters from the arrival time only, Bull. St. Louis Univ. 8 (1912) 60–71.
[14] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., 2001.
[15] S. Jin, H. Wu, X. Yang, Gaussian beam methods for the Schrödinger equation in the semi-classical regime: Lagrangian and Eulerian formulations, Commun. Math. Sci. 6 (4) (2008) 995–1020.
[16] L.V. Kantorovich, G.S. Rubinshtein, On a space of totally additive functions, Vestn. Leningr. Univ. 13 (7) (1958) 52–59.
[17] Y.H. Kim, Q.Y. Liu, J. Tromp, Adjoint centroid-moment tensor inversions, Geophys. J. Int. 186 (2011) 264–278.
[18] D. Komatitsch, J. Tromp, A perfectly matched layer absorbing boundary condition for the second-order seismic wave equation, Geophys. J. Int. 154 (2003) 146–153.
[19] W.H.K. Lee, S.W. Stewart, Principles and Applications of Microearthquake Networks, Academic Press, 1981.
[20] K. Levenberg, A method for the solution of certain problems in least square, Q. Appl. Math. 2 (1944) 164–168.
[21] J.S. Li, D.H. Yang, H. Wu, X. Ma, A low-dispersive method using the high-order stereo-modelling operator for solving 2-D wave equations, Geophys. J. Int. 210 (2017) 1938–1964.
[22] Q.Y. Liu, J. Polet, D. Komatitsch, J. Tromp, Spectral-element moment tensor inversion for earthquakes in southern California, Bull. Seismol. Soc. Am. 94 (5) (2004) 1748–1761.
[23] R. Madariaga, Seismic source theory, in: S. Gerald (Ed.), Treatise on Geophysics, 2nd ed., Elsevier B.V., 2015, pp. 51–71.
[24] K. Madsen, H.B. Nielsen, O. Tingleff, Methods for Non-Linear Least Squares Problems, 2nd ed., Informatics and Mathematical Modelling, Technical University of Denmark, 2004.
[25] D. Marquardt, An algorithm for least squares estimation on nonlinear parameters, SIAM J. Appl. Math. 11 (1963) 431–441.
[26] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, J. Virieux, Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion, Geophys. J. Int. 205 (2016) 345–377.
[27] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, J. Virieux, An optimal transport approach for seismic tomography: application to 3D full waveform inversion, Inverse Probl. 32 (2016) 115008.
[28] J. Milne, Earthquakes and Other Earth Movements, 1886, Appleton, New York.
[29] G. Monge, Mémoire sur la théorie des déblais et de remblais, in: Histoire de l'Académie royale des sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la mme année, 1781, pp. 666–704.
[30] A.F. Prugger, D.J. Gendzwill, Microearthquake location: a nonlinear approach that makes use of a simplex stepping procedure, Bull. Seismol. Soc. Am. 78 (1988) 799–815.
[31] N. Rawlinson, S. Pozgay, S. Fishwick, Seismic tomography: a window into deep Earth, Phys. Earth Planet. Inter. 178 (2010) 101–135.
[32] Y. Rubner, C. Tomasi, L.J. Guibas, A metric for distributions with applications to image databases, in: IEEE International Conference on Computer Vision, 1998, pp. 59–66.
[33] F. Santambrogio, Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling, Progress in Nonlinear Differential Equations and Their Applications, Birkhäuser, 2015.
[34] C. Satriano, A. Lomax, A. Zollo, Real-time evolutionary earthquake location for seismic early warning, Bull. Seismol. Soc. Am. 98 (3) (2008) 1482–1494.
[35] C.H. Thurber, Nonlinear earthquake location: theory and examples, Bull. Seismol. Soc. Am. 75 (3) (1985) 779–790.
[36] C.H. Thurber, Earthquake, location techniques, in: H.K. Gupta (Ed.), Encyclopedia of Earth Sciences Series, Springer, 2014, pp. 201–207.
[37] P. Tong, D.H. Yang, Q.Y. Liu, X. Yang, J. Harris, Acoustic wave-equation-based earthquake location, Geophys. J. Int. 205 (1) (2016) 464–478.
[38] P. Tong, D.H. Yang, D. Li, Q.Y. Liu, Time-evolving seismic tomography: the method and its application to the 1989 Loma Prieta and 2014 South Napa earthquake area, California, Geophys. Res. Lett. 44 (7) (2017) 3165–3175.
[39] P. Tong, D.P. Zhao, D.H. Yang, Tomography of the 1995 Kobe earthquake area: comparison of finite-frequency and ray approaches, Geophys. J. Int. 187 (2011) 278–302.

[40] C. Villani, Topics in Optimal Transportation, Graduate Studies in Mathematics, American Mathematical Society, 2003.
[41] C. Villani, Optimal Transport: Old and New, Springer Science & Business Media, 2008.
[42] F. Waldhauser, W.L. Ellsworth, A double-difference earthquake location algorithm: method and application to the northern Hayward Fault, California, Bull. Seismol. Soc. Am. 90 (6) (2000) 1353–1368.
[43] X. Wen, High order numerical quadratures to one dimensional delta function integrals, SIAM J. Sci. Comput. 30 (4) (2008) 1825–1846.
[44] H. Wu, J. Chen, X.Y. Huang, D.H. Yang, A new earthquake location method based on the waveform inversion, Commun. Comput. Phys. 23 (1) (2018) 118–141.
[45] H. Wu, J. Chen, H. Jing, P. Tong, D.H. Yang, The auxiliary function method for waveform based earthquake location, arXiv:1706.05551, 2017.
[46] H. Wu, X. Yang, Eulerian Gaussian beam method for high frequency wave propagation in the reduced momentum space, Wave Motion 50 (6) (2013) 1036–1049.
[47] Y.N. Yang, B. Engquist, J.Z. Sun, B.D. Froese, Application of optimal transport and the quadratic Wasserstein metric to Full-Waveform-Inversion, Geophysics 83 (1) (2018) R43–R62.