Convexity of the quadratic Wasserstein metric as a misfit function for full waveform inversion

Yunan Yang, Björn Engquist and Junzhe Sun, The University of Texas at Austin

SUMMARY

We analyze the properties of Wasserstein metric, a new misfit function for full waveform inversion (FWI) and prove such properties as convexity in different aspects. Considering the observed data and predicted data as two density functions, the quadratic Wasserstein metric corresponds to the optimal cost of rearranging one function into the other with a cost function that is quadratic in distance. In other words, we match the observed data and the predicted data by the optimal map which takes the information geometry of the data sets into consideration. The inversion follows the normal scheme of FWI as a PDE-constrained optimization. The velocity model can be updated using a gradient- based optimization with the new adjoint source.

INTRODUCTION

Full waveform inversion (FWI) was initially developed almost three decades ago in an attempt to obtain from seismic data quantitative information about subsurface properties on a detailed scale (Lailly, 1983; Tarantola, 1984). Over the last few years, there have been many encouraging results employing FWI in seismic processing of marine and land data (Virieux and Operto, 2009). FWI iteratively updates an estimated subsurface model and computes corresponding synthetic data to reduce the difference (the data misfit) between the synthetic and recorded data.

The objective of FWI is to match the synthetic and recorded data in a comprehensive way such that all information in waveforms is accounted for in the data misfit. The unknown wave velocity is determined by minimizing the mismatch $d(f,g)$ between the predicted data and the measured data.

FWI has the potential to generate high resolution quantitative models of the subsurface, but suffers from the ill-posedness of the inverse problem. This issue can be handled by considering multiple data components ranging from low to high frequency or by adding regularization terms (Virieux and Operto, 2009).

The least squares norm (L_2 norm) is the most used misfit function in FWI, but suffers from cycle skipping and many local minima, as well as sensitivity to noise. Other norms were studied in literature. The least-absolute-values norm (L_1 norm), Huber criterion and hybrid L_1/L_2 norm show some improvement compared with the least squares norm (Brossier et al., 2010). These misfit functions all follow the same path of dealing with the predicted data and observed data independently without considering the information geometry between them.

The difference between predicted velocity model and the true model is the reason of having misfit, which is the information we use to update the velocity model. This motivates us to take quite a different view on the predicted data and synthetic data by considering a "map" connecting them (Ma and Hale, 2013).

Following the idea that changes in velocity cause shift or "transport" in the arrival time of the signal, in this paper we study using the quadratic Wasserstein metric for the misfit function (Engquist and Froese, 2014).The Wasserstein metric is a concept from the optimal transportation in mathematics (Villani, 2003). Vividly, here we treat our data sets as density functions of two probability distributions, which can be imagined as the distributions for two piles of sand of equal mass. Given certain cost function, different shipping plan could generate different costs, and the plan corresponding the lowest cost is optimal map and the cost is our misfit.

In this paper, we focus on the quadratic Wasserstein metric (W_2) as a follow-up of the paper by Engquist and Froese (2014). Métivier et al. (2016) adapted a similar misfit function, but with a linear cost function. However, the optimal map is not unique for the linear cost function. Therefore, the optimization process is not rigorously equivalent to correcting the map to be identity, which may lost the convexity the quadratic metric has.

In this paper, we use mathematical theorems in optimal transport to prove the convexity of the quadratic Wasserstein metric with respect to shift, dilation, and partial amplitude change. We also prove its insensitivity to noise. We present a simple demonstration of the properties and inversion results of quadratic-Wasserstein based FWI using a gradient-based algorithm.

THEORY

The quadratic Wasserstein metric measures the distance between two distributions as the optimal cost of rearranging one distribution into the other. The mathematical definition of the distance between the distributions $f: X \to \mathbb{R}^+$ and $g: Y \to \mathbb{R}^+$ can be formulated as

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

where $\mathscr{M}$ is the set of all maps that rearrange the distribution f into g (Villani, 2003). The optimal transport formulation requires nonnegative distributions and equal total mass.

We are interested in computing the Wasserstein metric between two distributions f, g, which are supported on a rectangle region X. This can be accomplished via the solution of the Monge-Ampère equation (Engquist and Froese, 2014)

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

The squared Wasserstein metric is then given by

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

We solve the Monge-Ampère equation numerically using an almost-monotone finite difference method relying on a reformulation of the Monge-Ampère operator, which automatically enforces the convexity constraint (Froese, 2012).

The formula for Fréchet gradient of the squared Wasserstein metric with respect to the data f was introduced by Engquist et al. (2016). The gradient needed for the minimization is obtained through the composition $\nabla_f W_2^2 \nabla_v f$. As long as $\nabla_f W_2^2$ can be computed efficiently, techniques such as the adjoint state method can be used to efficiently construct the required gradient (Plessix, 2006).

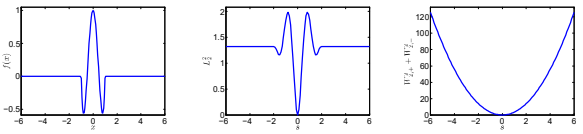
The adjoint source term can be expressed as

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

Notice that once the Monge-Ampère equation (2) has been solved, this gradient is easy to compute as it only requires the inversion of a single matrix that is already being inverted as a part of the solution of the Monge-Ampère equation.

PROPERTIES

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



(a) (b) (c)

Figure 1: (a) A synthetic wavelet profile $f(x)$. (b) The distance between $f(x)$ and $g(x) = f(x-s)$ measured by $L_2^2(f,g)$. (c) Same difference measured by $W_2^2(f^+,g^+) + W_2^2(f^-,g^-)$. The figure is reproduced from (Engquist and Froese, 2014).

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

Convexity with respect to shift

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

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

Proof of Theorem 1. By our original assumption, the optimal map between two measures with density functions f and g is T. We will show that the new joint measure $\pi_s = (Id \times T_s)\#\mu_s$ is cyclically monotone. This is based on two lemmas from Villani (2003) on the equivalence of optimality and cyclical monotonicity under the condition that μ does not give mass to small sets.

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

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

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

By the uniqueness of monotone measure-preserving optimal maps between two distributions (McCann, 1995), we assert that $T_s(x) = T(x - s\eta)$ is the optimal map corresponding to the shifted function $f(x - s\eta)$.

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

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

If we consider $W_2^2(f_s, g)$ as a function of the shift parameter s, the convexity with respect to s is evident from the last line. □

Convexity with respect to dilation

Next, we give the result on the formulation of optimal map for a change caused by dilation:

Theorem 2 (Optimal map for dilation)**.** *Assume $g(y)$ is a density function of finite second moment and $f(x) = \det(A)g(Ax)$, where A is a symmetric positive definite matrix. Then the optimal transport map to rearrange $f(x)$ into $g(y)$ is $T(x) = Ax$.*

Convexity is a separate question as it depends on the parameterisation. One special case of dilation occurs when A is a diagonal matrix. The following theorem is a natural consequence of the definition of the Wasserstein metric and Theorem 2.

Theorem 3 (Convexity with respect to dilation)**.** *Assume $g(y)$ is a density function of finite second moment and $f_\lambda(x) = \frac{1}{\lambda^n} g(\frac{x}{\lambda})$. Then the optimal map between f and g is $T(x) = \frac{x}{\lambda}$. The Wasserstein metric $W_2^2(f_\lambda, g)$ is $(1-\lambda)^2 \int y^2 g(y)dy$, a convex function of λ. More generally, if the dilation matrix $A = diag(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n})$ and $f(x) = \det(A)g(Ax)$, the Wasserstein metric $W_2^2(f,g)$ is convex with respect to $\lambda_1, \ldots, \lambda_n$.*

Remark 1. If both dilation and shift are present, the Wasserstein metric will be convex with respect to each of the corresponding parameters.

Convexity with respect to partial amplitude change

Now we consider density functions f_α and g on a domain $\Omega = \Omega_1 \cup \Omega_2$ with $\Omega_1 \cap \Omega_2 = \emptyset$. Assume function f_α depends on g as follows:

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

Here α is related to the size of the "perturbation". Obviously, $f_0(x) = g(x)$ for any $x \in \Omega$. The rescaling parameter γ ensures both distributions have the same mass $\int_\Omega f_\alpha(x)dx = \int_\Omega g(x)dx$, which is necessary to evaluate the Wasserstein metric.

Theorem 4 (Convexity with respect to partial amplitude change)**.** *With the density functions f_α and g defined as above, the Wasserstein metric $W_2^2(f_\alpha, g)$ is a convex function of the parameter α.*

Proof. Choose any α_1, α_2 such that f_{α_1} and f_{α_2} are nonnegative, $s \in [0,1]$, and let h be an arbitrary density function. From convexity of the Monge-Kantorovich minimization problem Villani (2003), we have

$$W_2^2(h, sf_{\alpha_1} + (1-s)f_{\alpha_2}) \le sW_2^2(h, f_{\alpha_1}) + (1-s)W_2^2(h, f_{\alpha_2}). \tag{4}$$

We can calculate

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

Thus we can rewrite Equation (4) as

$$W_2^2(h, f_{s\alpha_1+(1-s)\alpha_2}) \le sW_2^2(h, f_{\alpha_1}) + (1-s)W_2^2(h, f_{\alpha_2}) \tag{5}$$

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

NUMERICAL EXAMPLES

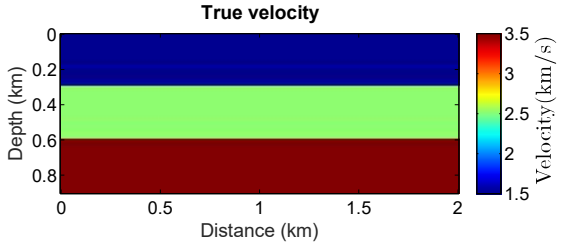


Figure 2: The three layer model used for the comparison between misfit functions.

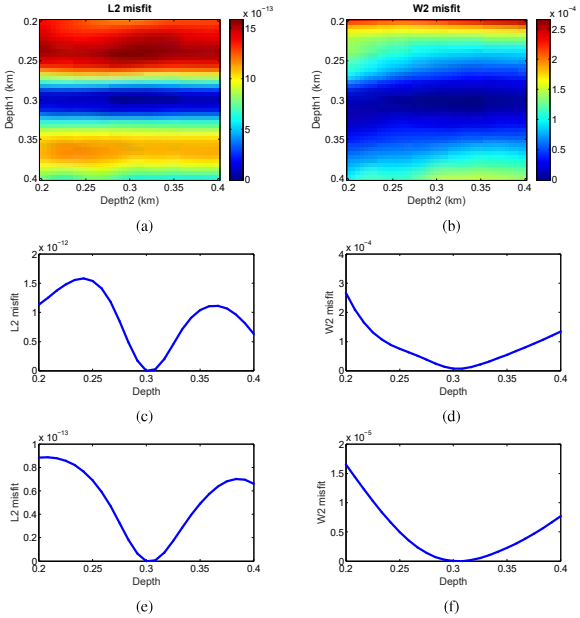


Figure 3: (a)(b) shows 2D misfit functions. (c)(d) shows the cross sections of misfits for the first layer. (e)(f) shows the cross sections of misfits for the second layer. The first column is L_2 norm result. The second column is W_2 misfit result.

We are inspired by an example in (Zhu and Fomel, 2016)to compare the behavior of misfit functions based on L_2 waveform differences and the quadratic Wasserstein distance. A simple 2D three-layer model Figure is used here. The model parameters are the depths for the first and second layer. By changing these two model parameters, we are able to compare the behavior of the misfit functions and draw the conclusion that L_2 waveform misfit involves local minima while the quadratic Wasserstein distance behaves well and avoids cycle skipping.

Marmousi model

To further demonstrate the properties of the Wasserstein metric, we apply it on one part of the well studied Marmousi model as Figure 4. The value of the velocity ranges from 1.6 km/s to 5.5 km/s.

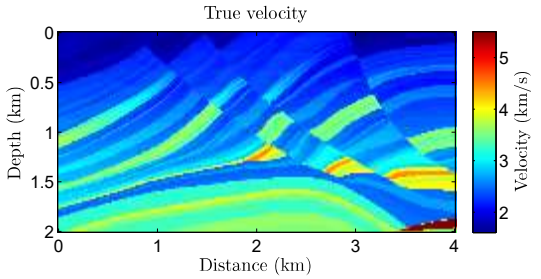


Figure 4: The true velocity model: part of the Marmousi model velocity field.

To illustrate convexity for shift and dilation from the previous section, we compare the wavefield $f(v)$ with wavefield $f(v_s)$ and $f(v_\lambda)$, generated by the true velocity v, the shifted velocity $v_s = v + s$ and the dilated velocity $v_\lambda = \lambda v$ respectively. We consider the W_2 misfit as a function of the shift s as in Figure 5a and a function of dilation parameter λ as Figure 5b shows, the convexity is obvious.

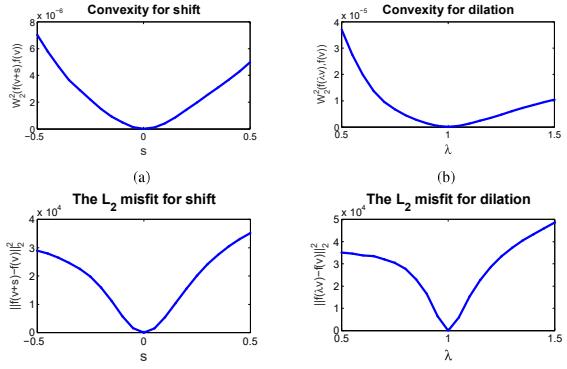


Figure 5: (a)(c) The W_2 and L_2 misfit between $f(v)$ and $f(v+s)$ (b)(d) The W_2 and L_2 misfit between $f(v)$ and $f(\lambda v)$.

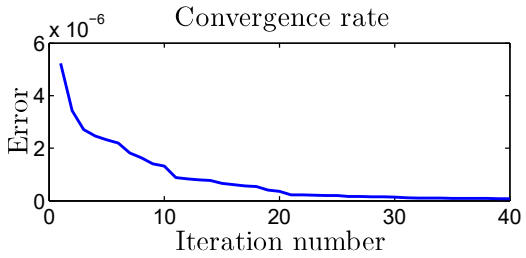


Figure 6: The convergence rate for the FWI with W_2 as misfit function.

The goal of FWI is to minimise the Wasserstein metric between computed data $f(v)$ and true data $f(v_0)$, where v_0 is the true velocity Marmousi model in Figure 4 and v is a strongly smoothed velocity. We follow the popular scheme using nonlinear conjugate gradient optimization. After 40 iterations we get the misfit within tolerance shown in Figure 6. Figure 7 shows the data residual for L_2 misfit and W_2 misfit respectively. It is clear to see the residual become much smaller after inversions.

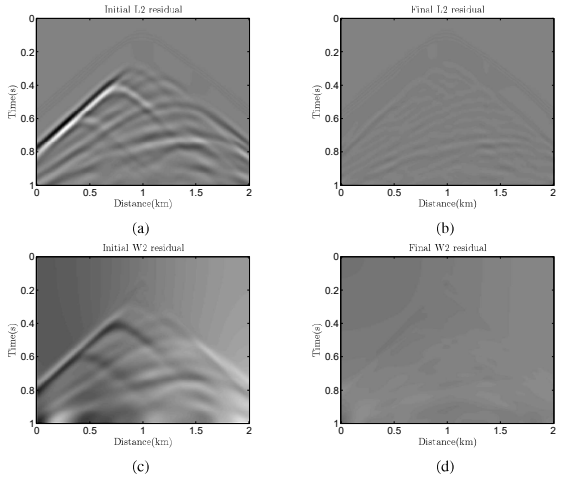


Figure 7: (a) The initial L_2 residual (b) The L_2 residual after convergence (c) The initial W_2 residual (d) The W_2 residual after convergence

CONCLUSION

In this paper, we study the ability of a new misfit function, the Wasserstein metric, to deal with cycle skipping and local minima problems in FWI. The quadratic Wasserstein metric is a true distance in optimal transportation which connects the observed data and predicted data by an optimal map. Based on its mathematical properties, we demonstrate the convexity with respect to shift, dilation and partial amplitude change with rescalling, which correspond to fundamental kinds of changes in wavefield due to incorrect velocity. In the end, we illustrate the advantage of W_2 over the traditional least squares norm. We also successfully applied the optimal transportation based misfit function on a 2D time-domain FWI with a Marmousi model. Numerical examples have indicated that the quadratic Wasserstein metric has the desirable properties of an optimal misfit function in seismic inversion.

ACKNOWLEDGMENTS

We thank Sergey Fomel and Zhiguang Xue for helpful discussions, and thank the sponsors of the Texas Consortium for Computational Seismology (TCCS) for financial support. We also thank Brittany Froese for discussions and her generosity of sharing the code.

EDITED REFERENCES

Note: This reference list is a copyedited version of the reference list submitted by the author. Reference lists for the 2016 SEG Technical Program Expanded Abstracts have been copyedited so that references provided with the online metadata for each paper will achieve a high degree of linking to cited sources that appear on the Web.

REFERENCES

Brossier, R., S. Operto, and J. Virieux, 2010, Which data residual norm for robust elastic frequency-domain full waveform inversion?: Geophysics, **75**, no. 3, R37–R46, http://dx.doi.org/10.1190/1.3379323.

Engquist, B., and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Sciences, 12, 978–988, http://dx.doi.org/10.4310/CMS.2014.v12.n5.a7

Engquist, B., B. D. Froese, and Y. Yang, 2016, Optimal transport for seismic full waveform inversion: arXiv preprint arXiv: 1602.01540.

Froese, B. D., 2012, A numerical method for the elliptic Monge-Amp`ere equation with transport boundary conditions: SIAM Journal on Scientific Computing, **34**, A1432–A1459, http://dx.doi.org/10.1137/110822372.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Conference on inverse scattering: Theory and application: Society for Industrial and Applied Mathematics, 206–220.

Ma, Y., and D. Hale, 2013, Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion: Geophysics, **78**, no. 6, R223–R233, http://dx.doi.org/10.1190/geo2013-0004.1.

McCann, R. J., 1995, Existence and uniqueness of monotone measure-preserving maps: Duke Mathematical Journal, **80**, 309–323, http://dx.doi.org/10.1215/S0012-7094-95-08013-2.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Journal International, **205**, 345–377, http://dx.doi.org/10.1093/gji/ggw014.

Plessix, R.-E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503, http://dx.doi.org/10.1111/j.1365-246X.2006.02978.x.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266, http://dx.doi.org/10.1190/1.1441754.

Villani, C., 2003, Topics in optimal transportation: American Mathematical Society, Vol 58, Graduate Studies in Mathematics.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, no. 6, WCC1–WCC26, http://dx.doi.org/10.1190/1.3238367.

Zhu, H., and S. Fomel, 2016, Building good starting models for full waveform inversion using adaptive using adaptive matching filter: Geophysics, submitted.