GEOPHYSICS, VOL. 83, NO. 1 (JANUARY-FEBRUARY 2018); P. A7–A12, 4 FIGS.
10.1190/GEO2017-0264.1

Analysis of optimal transport and related misfit functions in full-waveform inversion

Yunan Yang[1] and Björn Engquist[1]

ABSTRACT

Full-waveform inversion has evolved into a powerful computational tool in seismic imaging. New misfit functions for matching simulated and measured data have recently been introduced to avoid the traditional lack of convergence due to cycle skipping. We have introduced the Wasserstein distance from optimal transport for computing the misfit, and several groups are currently further developing this technique. We evaluate three essential observations of this new metric with implication for future development. One is the discovery that trace-by-trace comparison with the quadratic Wasserstein metric works remarkably well together with the adjoint-state method. Another is the close connection between optimal transport-based misfits and integrated techniques with normalization as, for example, the normalized integration method. Finally, we study the convexity with respect to selected model parameters for different normalizations and remark on the effect of normalization on the convergence of the adjoint-state method.

INTRODUCTION

Full-waveform inversion (FWI) is a data-driven method in seismology to obtain high-resolution subsurface properties by minimizing the difference between observed and synthetic seismic waveforms (Virieux et al., 2017). In the past three decades, the least-squares norm (L^2) has been widely used as a misfit function (Tarantola and Valette, 1982; Lailly, 1983), which is known to suffer from cycle-skipping issues with local minimum trapping and sensitivity to noise (Symes, 2008; Virieux and Operto, 2009). Other misfit functions proposed in literature (Brossier et al., 2010; Bozdag et al., 2011), include the Huber norm (Ha et al., 2009), filter-based misfit functions (Warner and Guasch, 2014; Zhu and Fomel, 2016), seismic envelop inversion (Luo and Wu, 2015), etc. The lower frequency components have a wider basin of attraction with the least-squares norm as the misfit function. Several hierarchical methods that invert from low frequencies to higher frequencies have been proposed in literature to mitigate the cycle skipping of the inverse problem (Kolb et al., 1986; Pratt and Worthington, 1990; Bunks et al., 1995; Weglein et al., 2003; Sirgue and Pratt, 2004).

A recently introduced class of misfit functions is optimal transport related (Engquist and Froese, 2014; Engquist et al., 2016; Métivier et al., 2016a, 2016b; Yang et al., 2016). As useful tools from the theory of optimal transport, the quadratic Wasserstein metric (W_2) computes the minimal cost of rearranging one distribution into another with a quadratic cost function, and the 1-Wasserstein metric (W_1) using the absolute value cost function. Although the L^2 misfit function measures the difference in amplitude locally, the optimal transport-based methods compare the observed and simulated data globally and thus include phase information. Researchers started to combine these two parts in velocity analysis a long time ago. The differential semblance optimization (Symes and Carazzone, 1991) exploits the phase and amplitude information of the reflections. Tomographic FWI (Biondi and Almomin, 2012) also has the global convergence characteristics of wave-equation migration velocity analysis.

The filter-based techniques are also of a global nature. First, a filter is designed to minimize the L^2 difference between the filtered simulated data and the observed data. The misfit is then a measure of how much the filter deviates from the identity. As we will see in the W_2 technique, this is done in one step in which the optimal map directly determines the mapping of the simulated data. The mapping in W_2 is general and does not need to have the form of a convolution filter.

In this paper, we will also discuss the integral wavefields misfit functional (Huang et al., 2014) and the normalized integration method (NIM) (Liu et al., 2012). If we consider that the data are properly rescaled, the misfit of NIM is the norm of Sobolev space H^{-1} in mathematics. The connection between W_2 and H^{-1} is not obvious from the optimal transport definition, but it is clear from its equivalent formulation and the 1D closed solution formula. We

Manuscript received by the Editor 29 April 2017; revised manuscript received 28 September 2017; published ahead of production 12 October 2017; published online 29 November 2017.
[1]The University of Texas at Austin, Department of Mathematics, Austin, Texas, USA. E-mail: yunanyang@math.utexas.edu; engquist@ices.utexas.edu.

shall also see that this is valid in higher dimensions even if there is no explicit solution formula.

THEORY

FWI is a PDE-constrained optimization problem, minimizing the data misfit $d(f,g)$ by updating the model parameter m, i.e.,

$$m^\star = \underset{m}{\operatorname{argmin}}\, d(f(x_r,t;m), g(x_r,t)), \tag{1}$$

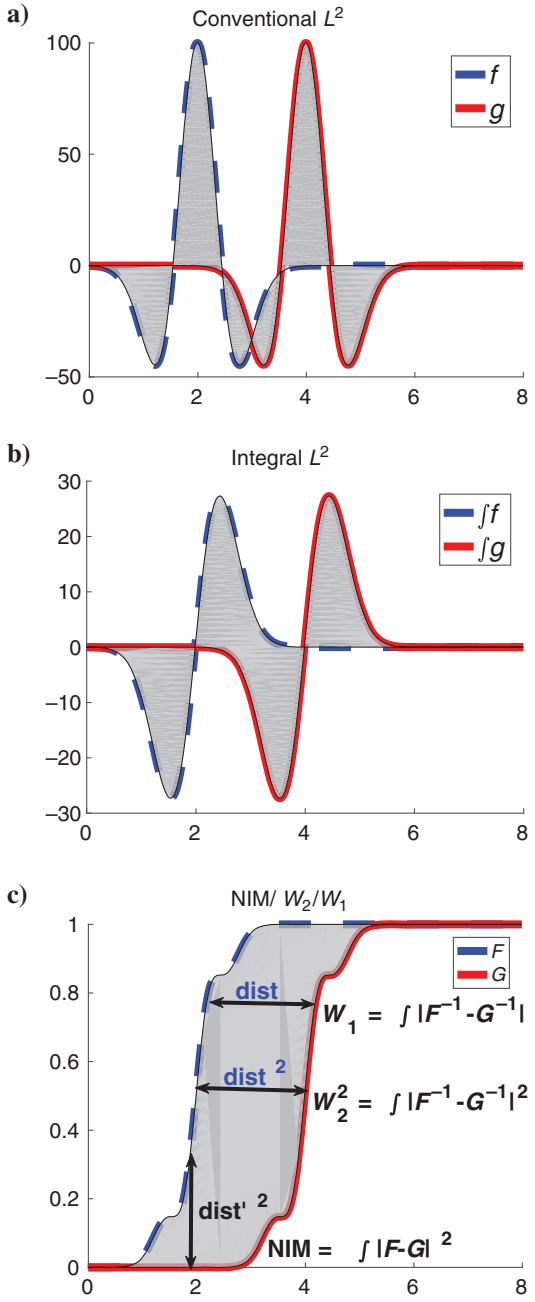

Figure 1. The shaded areas represent the mismatch each misfit function considers. (a) L^2: $\int(f-g)^2dt$. (b) Integral wavefield method: $\int(\int f-\int g)^2dt$. After data normalization, (c) NIM measures $\int(F-G)^2dt$, whereas W_2 considers $\int(F^{-1}-G^{-1})^2dt$ and W_1 considers $\int|F^{-1}-G^{-1}|dt$.

where g is the observed data, f is the simulated data, x_r are the receiver locations, and m is the model parameter. We get the modeled data $f(x,t;m)$ by solving a wave equation numerically in the space and time domain.

The generalized least-squares functional is a weighted sum of the squared errors and hence a generalized version of the standard least-squares misfit function. The formulation is

$$J_1(m) = \sum_r \int |W(f(x_r,t;m)) - W(g(x_r,t))|^2 dt, \tag{2}$$

where W is an operator. In the conventional L^2 misfit, $W=I$, the identity operator.

The integral wavefields misfit functional is a generalized least-squares functional applied on FWI with weighting operator $W(u) = \int_0^t u(x,\tau)d\tau$. The objective function is defined as

$$J_2(m) = \sum_r \int \left| \int_0^t f(x_r,\tau;m)d\tau - \int_0^t g(x_r,\tau)d\tau \right|^2 dt. \tag{3}$$

NIM is another generalized least-squares functional, similar to the integral wavefield misfit functional.

The difference from equation 3 is that NIM first makes data positive before the integration. The objective function is

$$J_3(m) = \sum_r \int |Q(f(x_r,t;m)) - Q(g(x_r,t))|^2 dt, \tag{4}$$

where Q is the transformation of the wavefield u, defined as

$$Q(u)(x_r,t) = \frac{\int_0^t N(u)(x_r,\tau)d\tau}{\int_0^T N(u)(x_r,\tau)d\tau}. \tag{5}$$

The operator N is included to make the data nonnegative. Three common choices are $N_1(u)=|u|$, $N_2(u)=u^2$, and $N_3=E(u)$, which correspond to the absolute value, the square, and the envelope of the signal (Liu et al., 2012).

Despite the fact that both methods are measuring the L^2 misfit, there are three different features in NIM compared with conventional FWI. Data sets are normalized to be nonnegative, mass balanced, and integrated in time. The first two are exactly the prerequisite of optimal transport-based misfit functions, i.e., the Wasserstein metrics.

Optimal transport refers to the problem of seeking the minimal cost required to transport mass of one distribution into another given a cost function, e.g., $|x-y|^p$. The mathematical definition of the distance between the distributions $f: X \to \mathbb{R}^+$ and $g: Y \to \mathbb{R}^+$ can then be formulated as

$$W_p^p(f,g) = \inf_{T_{f,g}\in\mathcal{M}} \int_X |x - T_{f,g}(x)|^p f(x)dx, \tag{6}$$

where p is typically one or two, $\mathcal{M}$ is the set of all maps $T_{f,g}$ that rearrange the distribution f into g (Villani, 2003).

The optimal transport formulation requires nonnegative distributions and equal total masses, $\int f(x)dx = \int g(x)dx$, which are not natural for seismic signals. Therefore, proper data normalization is required before inversion. Data sets f and g can be rescaled

to be nonnegative with values in the range $[0, 1]$, and to have equal mass. This step is the same as the one in equation 5 in NIM.

In Yang et al. (2016), two ways of using W_2 in FWI were proposed. One can either compute the misfit globally by solving a 2D optimal transport problem or compare data trace-by-trace with the 1D explicit formula. Here, we mainly focus on the 1D technique:

$$J_4(m) = \sum_{r=1}^{R} W_2^2(f(x_r, t; m), g(x_r, t)), \tag{7}$$

where R is the total number of traces. Mathematically, it is the W_2 metric in the time domain and the L^2 norm in the spatial domain.

PROPERTIES

Next, we discuss integration and positivity, two important features of optimal transport by comparing the misfit functions mentioned above. We will regard f and g as the synthetic and observed data from one single trace as a 1D problem.

Relations among misfit functions

Conventional FWI measures the L^2 norm difference $\int |f(t) - g(t)|^2 dt$, indicated by the shaded part in Figure 1a. The integral wavefield misfit functional first integrate f and g in time, and then it measures their L^2 misfit (equation 3). The integrated wavefields have the higher frequencies reduced compared with the lower ones. The reduced higher frequency components (in Figure 1b) can properly explain the improvement in inversion (Huang et al., 2014). It also motivates searching for new normalization functions that have better convexity properties.

With a proper normalization method, it is possible to scale the data to have nonnegativity and mass balance. This step is essential for NIM and W_2. We are able to solve the 1D optimal transport problem exactly, and the optimal map is the unique monotone rearrangement of the density f into g (Villani, 2003). To compute the quadratic Wasserstein metric, we need the cumulative distribution functions $F(t) = \int_0^t f(\tau) d\tau$ and $G(t) = \int_0^t g(\tau) d\tau$ and their inverses F^{-1} and G^{-1}. The explicit formula for the 1D Wasserstein metric is $W_p^p(f, g) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|^p dx$.

The interesting fact is that W_2 computes the L^2 misfit between F^{-1} and G^{-1} (Figure 1c), whereas the objective function of NIM measures the L^2 misfit between F and G, i.e., $\int_0^T |F(t) - G(t)|^2 dt$ (Figure 1c). The latter is identical to the mathematical norm of the Sobolev space H^{-1}, $\|f - g\|_{H^{-1}}^2$, given f and g are nonnegative and sharing equal mass.

The adjoint source of the integral wavefield misfit functional as well as some earlier works enhances the low frequency already present in the data. The Fréchet derivative of trace-by-trace W_2 has a lower frequency outside the data bandwidth especially when f and g are far apart. This is because optimal transport considers not only amplitude differences but also phase shifts (generating low frequencies in the adjoint source). The envelope inversion (Luo and Wu, 2015) has a similar property. Once the cycle-skipping problem is solved and the observed and simulated data are close, the W_2 norm is more similar to the point-by-point misfit functions such as L^1 and L^2.

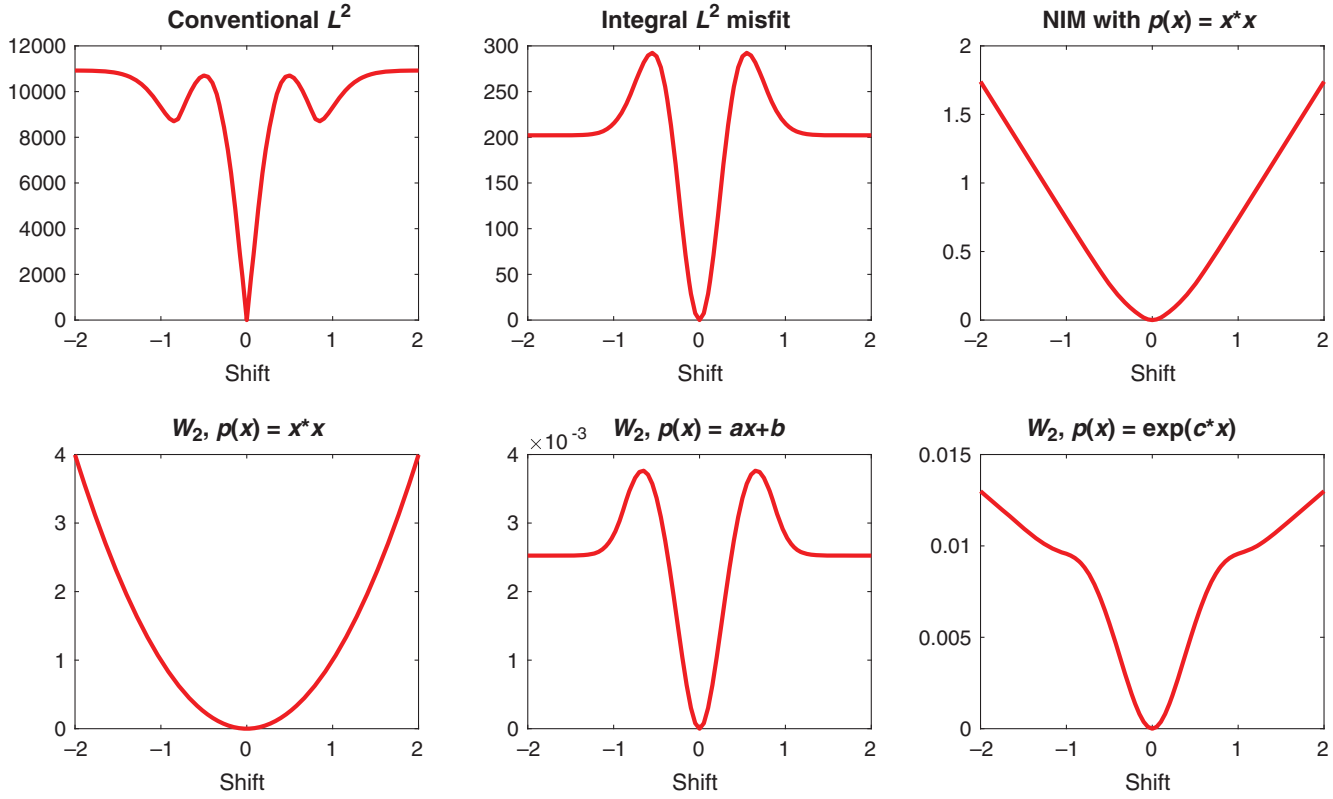

Figure 2. The misfit between $f(x)$ and $f(x - s)$ by six different misfit functions. The first row shows conventional L^2 (left), integral wavefield method (middle) and NIM with $P(f) = f^2$ (right). The second row shows the W_2 misfit with different normalization methods: $P(f) = f^2$ (left), $a \cdot f + b$ (middle) and $\exp(c \cdot f)$ (right).

Mathematical connection between H^{-1} norm and W_2 norm

In the general case, f and g are the synthetic and observed data in higher dimensions, satisfying nonnegativity and conservation of mass. To compute the quadratic Wasserstein metric, we solve the following Monge-Ampère equation (Brenier, 1991):

$$\det(D^2u(x)) = f(x)/g(\nabla u(x)). \tag{8}$$

If f and g are close enough and $g = (1 + \epsilon h + O(\epsilon^2))f$, where $\int h(x)f(x)dx = 0$, we can linearize (equation 8) and also derive an approximation of the quadratic Wasserstein metric between f and g as a weighted H^{-1} norm (Villani, 2003, 126–127).

The dynamical characterization of the Wasserstein metric proposed by Benamou and Brenier (2000) gives insights to consider that H^{-1} and W_2 belong to the same class of measures. One can refer to Dolbeault et al. (2009) for more theoretical details. The adjoint sources are, however, different for W_2 and NIM.

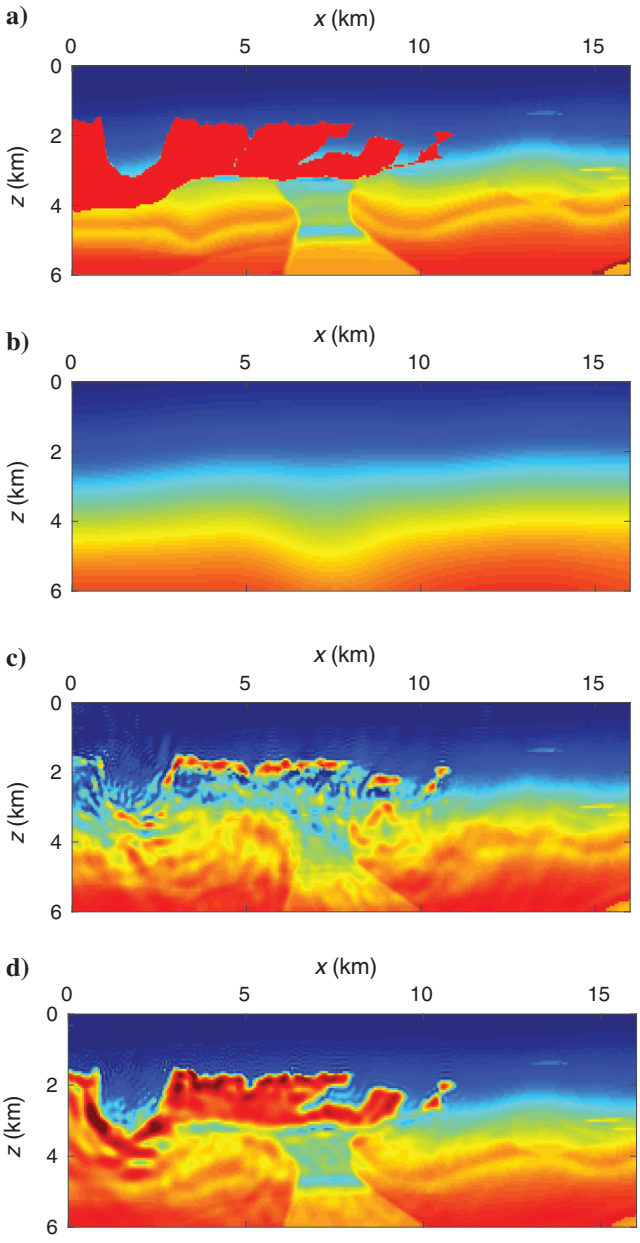

Figure 3. (a) True model velocity, (b) initial velocity, (c) inversion result using L^2, and (d) inversion result using trace-by-trace W_2 with normalization $P_2(f) = af + b$.

How to normalize the data: Convexity versus convergence

As demonstrated by Engquist et al. (2016), the squared Wasserstein metric has several properties that make it attractive as a choice of misfit function. One highly desirable feature is its convexity with respect to several parameterizations. However, the convexity of W_2 highly depends on the data normalization method to satisfy positivity and mass balance.

The curves in the second row of Figure 2 are the squared W_2 distance with different scaling functions: $P_1(f) = f^2$, $P_2(f) = a \cdot f + b$, and $P_3(f) = \exp(c \cdot f)$.

Theoretically, P_1 gives perfect convexity with respect to simple shifts but in most large-scale simulations with adjoint-state method P_2 has much better convergence properties. From Taylor expansion we can see that P_3 is very close to P_2 when c is small. Because P_2 and P_3 are similar also in larger scale FWI we have chosen to focus on the simpler P_2 in this paper. There are some reasons for P_2 to be successful even if Figure 2 indicates differently. One is that the situation in higher dimension is different.

The most important one is that optimal transport with linear normalization does not distort the shared events in data sets. It would also map the missing events in the synthetic data accurately if seismic data have mean-zero property.

NUMERICAL EXAMPLE

In this section, we use a part of the BP 2004 benchmark velocity model (Billette and Brandsberg-Dahl, 2005) (Figure 3a) and a highly smoothed initial model without the upper salt part (Figure 3b) to do inversion with W_2 and L^2 norm, respectively. A fixed-spread surface acquisition is used, involving 11 shots located every 1.6 km on top. A Ricker wavelet centered on 5 Hz is used to generate the synthetic data with a band-pass filter only keeping the 3–9 Hz components. We stopped the inversion after 300 L-BFGS iterations.

For W_2 we normalized the data with function $p_2(f) = a \cdot f + b$ to satisfy the nonnegativity and mass balance in optimal transport. Inversion with a trace-by-trace W_2 norm successfully constructed the shape of the salt bodies (Figure 3d), whereas FWI with the conventional L^2 failed to recover boundaries of the salt bodies as shown in Figure 3c.

Next, we repeat the previous experiment in a more realistic setting by adding correlated noise to the original observed data (Figure 4a). At each time grid, the noise $\tilde{r}(j) = 0.25 \cdot (r(j-1) + 2r(j) + r(j+1)) \cdot (1 + (g(j))/(\|g\|_\infty))$, where r is the mean-zero uniform iid noise and g is the clean observed data. The S/N is 5.98 dB.

After 185 iterations, the optimization converges to a velocity presented in Figure 4b. Compared with Figure 3d, the result has lower accuracy around the salt bottom due to the strong noise added in the target data (Figure 4a). However, we still can recover the salt body and upper boundaries reasonably well compared with Figure 3c when L^2 norm hardly generate meaningful results even with the

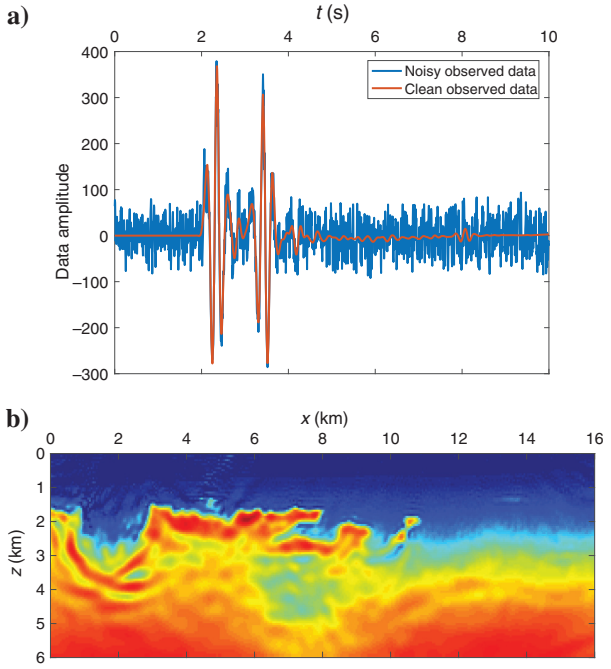

Figure 4. (a) Noisy and clean data of one trace and (b) inversion result with noisy target data.

clean observed data. The W_2 robustness to noise comes from cancellation locally (Engquist et al., 2016). This advantage is naturally reduced for signals highly correlated in time in the trace-by-trace technique.

CONCLUSION

In this paper, we mainly analyze the properties of trace-by-trace W_2 as a misfit function in FWI, which proved to be very successful in mitigating cycle skipping. Other misfit functions are considered for comparison. The W_2, the integral wavefields misfit functional and NIM in different ways incorporate the idea of integration. By itself, this cannot avoid local minima coming from the oscillatory data. One solution to reduce the risk of cycle skipping is to combine the integration with normalizing the signals to be nonnegative. This can "break" the periodicity.

The NIM and the quadratic Wasserstein metric include these ideas as essential steps and their convexity in the data domain and model domain become better. We have seen that the choice of normalization plays a significant role in convexity and convergence. The relation is not simple and needs to be further analyzed, which is seen from the examples of the successful linear normalization compared to squaring of the signals. The analysis of these misfit functions of FWI brings additional insights into the importance of seismic data preconditioning, which also can be seen in examples of large-scale FWI.

ACKNOWLEDGMENTS

We thank S. Fomel, J. Sun, L. Qiu, and Z. Xue for helpful discussions, and we thank the sponsors of the Texas Consortium for Computational Seismology (TCCS) for financial support. This work was also partially supported by NSF DMS-1620396.

REFERENCES

Benamou, J.-D., and Y. Brenier, 2000, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem: Numerische Mathematik, **84**, 375–393, doi: 10.1007/s002110050002.

Billette, F., and S. Brandsberg-Dahl, 2005, The 2004 BP velocity benchmark: Presented at the 67th Annual International Conference and Exhibition, EAGE.

Biondi, B., and A. Almomin, 2012, Tomographic full waveform inversion (TFWI) by combining full waveform inversion with wave-equation migration velocity analysis: 82nd Annual International Meeting, SEG, Expanded Abstracts, doi: 10.1190/segam2012-0275.1.

Bozdag, E., J. Trampert, and J. Tromp, 2011, Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements: Geophysical Journal International, **185**, 845–870, doi: 10.1111/j.1365-246X.2011.04970.x.

Brenier, Y., 1991, Polar factorization and monotone rearrangement of vector-valued functions: Communications on Pure and Applied Mathematics, **44**, 375–417, doi: 10.1002/(ISSN)1097-0312.

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, doi: 10.1190/1.3379323.

Bunks, C., F. M. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473, doi: 10.1190/1.1443880.

Dolbeault, J., B. Nazaret, and G. Savaré, 2009, A new class of transport distances between measures: Calculus of Variations and Partial Differential Equations, **34**, 193–231, doi: 10.1007/s00526-008-0182-5.

Engquist, B., and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Sciences, **12**, 979–988, doi: 10.4310/CMS.2014.v12.n5.a7.

Engquist, B., B. D. Froese, and Y. Yang, 2016, Optimal transport for seismic full waveform inversion: Communications in Mathematical Sciences, **14**, 2309–2330, doi: 10.4310/CMS.2016.v14.n8.a9.

Ha, T., W. Chung, and C. Shin, 2009, Waveform inversion using a back-propagation algorithm and a Huber function norm: Geophysics, **74**, no. 3, R15–R24, doi: 10.1190/1.3112572.

Huang, G., H. Wang, and H. Ren, 2014, Two new gradient precondition schemes for full waveform inversion: arXiv preprint arXiv:1406.1864.

Kolb, P., F. Collino, and P. Lailly, 1986, Pre-stack inversion of a 1-D medium: Proceedings of the IEEE, **74**, 498–508, doi: 10.1109/PROC.1986.13490.

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

Liu, J., H. Chauris, and H. Calandra, 2012, The normalized integration method — An alternative to full waveform inversion?: Presented at the 25th Symposium on the Application of Geophysics to Engineering & Environmental Problems.

Luo, J., and R.-S. Wu, 2015, Seismic envelope inversion: Reduction of local minima and noise resistance: Geophysical Prospecting, **63**, 597–614, doi: 10.1111/1365-2478.12208.

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

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016b, An optimal transport approach for seismic tomography: Application to 3D full waveform inversion: Inverse Problems, **32**, 115008, doi: 10.1088/0266-5611/32/11/115008.

Pratt, R. G., and M. Worthington, 1990, Inverse theory applied to multi-source cross-hole tomography. Part 1: Acoustic wave-equation method: Geophysical Prospecting, **38**, 287–310, doi: 10.1111/j.1365-2478.1990.tb01846.x.

Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248, doi: 10.1190/1.1649391.

Symes, W., and J. J. Carazzone, 1991, Velocity inversion by differential semblance optimization: Geophysics, **56**, 654–663, doi: 10.1190/1.1443082.

Symes, W. W., 2008, Migration velocity analysis and waveform inversion: Geophysical Prospecting, **56**, 765–790, doi: 10.1111/j.1365-2478.2008.00698.x.

Tarantola, A., and B. Valette, 1982, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics, **20**, 219–232, doi: 10.1029/RG020i002p00219.

Villani, C., 2003, Topics in optimal transportation: American Mathematical Society, Graduate studies in mathematics 58.

Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, 2014, 6. An introduction to full waveform inversion, *in* V. Grechka, and K. Wapenaar, eds., Encyclopedia of Exploration Geophysics: SEG, R1-1–R1-40.

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

Warner, M., and L. Guasch, 2014, Adaptive waveform inversion: Theory: 84th Annual International Meeting, SEG, Expanded Abstracts, 1089–1093.

Weglein, A. B., F. V. Araújo, P. M. Carvalho, R. H. Stolt, K. H. Matson, R. T. Coates, D. Corrigan, D. J. Foster, S. A. Shaw, and H. Zhang, 2003, Inverse scattering series and seismic exploration: Inverse Problems, **19**, R27–R83, doi: 10.1088/0266-5611/19/6/R01.

Yang, Y., B. Engquist, J. Sun, and B. D. Froese, 2016, Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion: arXiv preprint arXiv:1612.05075.

Zhu, H., and S. Fomel, 2016, Building good starting models for full-waveform inversion using adaptive matching filtering misfit: Geophysics, **81**, no. 5, U61–U72, doi: 10.1190/geo2015-0596.1.