On the use of optimal transport distances for a PDE-constrained optimization problem in seismic imaging

L. Métivier, A. Allain, R. Brossier, Q. Mérigot, E. Oudet, J. Virieux

Abstract Full waveform inversion is a PDE-constrained nonlinear least-squares problem dedicated to the estimation of the mechanical subsurface properties with high resolution. Since its introduction in the early 80s, a limitation of this method is related to the non-convexity of the misfit function which is minimized when the method is applied to the estimation of the subsurface wave velocities. Recently, the definition of an alternative misfit function based on an optimal transport distance has been proposed to mitigate this difficulty. In this study, we review the difficulties for exploiting standard optimal transport techniques for the comparison of seismic data. The main difficulty is related to the oscillatory nature of the seismic data, which requires to extend optimal transport to the transport of signed measures. We review three standard possible extensions relying on a decomposition of the data into its positive and negative part. We show how the two first might not be adapted for full waveform inversion and focus on the third one. We present a numerical strategy based on the dual formulation of a particular optimal transport distance yielding an efficient implementation. The interest of this approach is illustrated on the 2D benchmark Marmousi model.

L. Métivier
ISTerre/LJK, CNRS, Univ. Grenoble Alpes, France, e-mail: ludovic.metivier@univ-grenoble-alpes.fr

A. Allain
LJK, Univ. Grenoble Alpes, France, e-mail: aude.allain@imag.fr

R. Brossier
Univ. Grenoble Alpes, ISTerre, Grenoble, France, e-mail: romain.brossier@univ-grenoble-alpes.fr

Q. Mérigot
LMO, Univ. Paris Sud, France, e-mail: quentin@mgrt.fr

E. Oudet
LJK, Univ. Grenoble Alpes, France, e-mail: edouard.oudet@imag.fr

J. Virieux
Univ. Grenoble Alpes, ISTerre, Grenoble, France, e-mail: jean.virieux@univ-grenoble-alpes.fr

1 Full waveform inversion as a PDE-constrained nonlinear optimization problem

Full waveform inversion (FWI) is a high resolution seismic imaging technique which aims at reconstructing subsurface mechanical properties such as wave velocities, density, attenuation or anisotropy parameters, from the recording of seismic waves at the surface. Compared to conventional tomography strategies, based on the interpretation of arrival times only, FWI should exploit the totality of the seismic signal, which is expected to provide higher resolution estimates of the subsurface parameters, in the limit of half the shortest wavelength of the propagated signal following the theory of diffraction tomography [Devaney, 1984]. A recent review of FWI is proposed by [Virieux et al., 2017]. FWI is usually formulated as the minimization over the space of the subsurface parameters of the misfit between predicted and observed data. The predicted data is computed through the solution of partial differential equations (PDE) describing the seismic waves propagation. In the simplest settings, which we consider in this study, the acoustic approximation is adopted. Using this formalism, the problem is cast as the following PDE-constrained nonlinear optimization problem [Lailly, 1983, Tarantola, 1984]

$$\begin{cases} \min_{v_P} J(v_P) = g(d_{cal}, d_{obs}) + \alpha R(v_P),\ v_P(x) \in \mathscr{C}^p(\Omega),\ \ \Omega \subset \mathscr{R}^d \\ \dfrac{1}{\rho v_P(x)^2} \partial_{tt} u(x,t) - \ \text{div}\left(\dfrac{1}{\rho(x)} \nabla u(x,t) \right) = s(x,t),\ (x,t) \in \Omega \times [0,T], \\ d_{cal}(x_r,t) = H(u)(x_r,t),\ (x_r,t) \in \Gamma \times [0,T]. \end{cases} \tag{1}$$

In the system (1), the spatial domain Ω is a subset of $\mathscr{R}^d$, where $d = 2$ or $d = 3$, while Γ denotes a subset of the border $\partial\Omega$. The time interval is defined by $[0,T]$, where $T > 0$. The control variable is denoted by $v_P(x)$: this is the pressure wave (P-wave) velocity, which is supposed to be smooth up to a certain level of regularity $p \in \mathscr{N}$. The P-wave velocity is generally the main parameter to be reconstructed, even if the density $\rho(x)$ can also be included in the inverse problem yielding a so-called multi-parameter problem (see [Operto et al., 2013] for a review on multi-parameter FWI). The functional $J(v_P)$ measures the misfit between predicted data $d_{cal}(x_r,t)$ and observed data $d_{obs}(x_r,t)$ through a misfit measurement function g which is often taken as the least-squares norm

$$g(d_{cal}, d_{obs}) = \frac{1}{2} \| d_{cal} - d_{obs} \|_{L^2}^2. \tag{2}$$

It shall be noted that this least-squares distance measure is local: each sample of the observed data is compared with its synthetic counterpart at the same position in the data space, neglecting any information which could come from the neighboring samples. As a result, the least-squares distance is unable to detect shifted patterns between two datasets.

A regularization term $R(v_p)$, weighted by a positive coefficient α, is also generally added to the misfit measurement to reduce the null space of the underlying

inverse problem. Usual choices for $R(v_P)$ include prior information regularization, or penalization of the first-order spatial derivatives (Tikhonov regularization)

$$R(v_P) = \frac{1}{2}\|v_P - v_{P,0}\|_{L^2}^2, \;\; R(v_P) = \sum_{i=1}^{d} \frac{1}{2}\|\partial x_i v_P\|_{L^2}^2. \tag{3}$$

The calculated data $d_{cal}(x_r,t)$ is computed from the solution $u(x,t)$ of the acoustic wave equation through the observation operator $H(u)$. In practice, this observation operator simply extracts the value of the wavefield $u(x,t)$ at the receivers locations.

A Lagrangian function associated with the PDE-constrained problem (1) is

$$\begin{aligned} L(v_P, d_{cal}, u, \lambda_1, \lambda_2) &= g(d_{cal}, d_{obs}) + \alpha R(v_P) \\ &+ \int_{x_r \in \Gamma} \int_0^T (d_{cal}(x_r,t) - Hu(x_r,t))\, \lambda_2(x_r,t) dx_r dt \\ &+ \int_{x \in \Omega} \int_0^T \left(\frac{1}{\rho v_P^2} \partial_{tt} u(x,t) - \text{ div}\left(\frac{1}{\rho}\nabla u(x,t)\right) - s(x,t)\right) \lambda_1(x,t) dx dt \end{aligned} \tag{4}$$

First-order Karush-Kuhn-Tucker conditions give necessary conditions to characterize the solution of (1). They are obtained by canceling the first-order partial derivatives of the Lagrangian operator.

$$\begin{cases} -\dfrac{2}{\rho v_P^3} \displaystyle\int_0^T \partial_{tt} u(x,t) \lambda_1(x,t) dt + \alpha \nabla R(v_P) = 0 & (5) \\ d_{cal} = H(u) & (6) \\ \dfrac{1}{\rho v_P^2} \partial_{tt} u - \text{ div}\left(\dfrac{1}{\rho}\nabla u\right) = s & (7) \\ \lambda_2 = -\partial_{d_{cal}} g(d_{cal}, d_{obs}) & (8) \\ \partial_{tt}\lambda_1 - \rho v_P^2 \text{ div}\left(\dfrac{1}{\rho}\nabla \lambda\right) = -\partial_u H(u) \lambda_2 & (9) \end{cases}$$

Instead of solving the Karush-Kuhn-Tucker system iteratively through a Newton algorithm, a "reduced space" method is preferred [Nocedal and Wright, 2006] for efficiency. The misfit function $J(v_P)$ is minimized following iterative local optimization methods for smooth nonlinear functions, which rely on the ability to compute its gradient $\nabla J(v_P)$. This gradient is computed from the equation

$$\nabla J(v_P) = -\frac{2}{\rho v_P^3} \int_0^T \partial_{tt} \overline{u}(x,t) \overline{\lambda}_1(x,t) dt + \alpha \nabla R(v_P), \tag{10}$$

where fields $\overline{u}(x,t)$ and $\overline{\lambda}_1(x,t)$ are obtained through the solution of the equations from (6) to (9). In particular, using the L^2 norm for the definition of the misfit measurement function g yields the simple expression

$$\lambda_2 = -(d_{cal} - d_{obs}). \tag{11}$$

The reduced space method thus yields an efficient strategy to compute the gradient $\nabla J(v_P)$. This technique, also introduced as the adjoint state method within the optima control theory [Lions, 1968], has been known for a long time in seismic imaging [Chavent, 1971] and in weather forecasting [Le Dimet and Talagrand, 1986]. A review of the adjoint state method and its application in seismic imaging has been proposed by [Plessix, 2006].

Among different minimization strategies, the nonlinear conjugate gradient method, the quasi-Newton l-BFGS [Nocedal, 1980] or the truncated Newton approach [Nash, 2000] are used to solve the FWI problem (see [Métivier and Brossier, 2016] for a review of standard minimization algorithms used in FWI).

Since its introduction in the 80's, one of the main challenge for FWI is related to the non-convexity of the P-wave velocity reconstruction problem. For practical applications, the size of the discrete problem prevents the use of global or semi-global optimization strategies (Monte-Carlo or genetic algorithms for instance): in 2D the number of unknowns easily reaches $O(10^6)$, in 3D this number grows up to $O(10^9)$. The use of local optimization strategies thus requires to start the iterative process from a v_P model close enough from the solution, otherwise the method converges to a local minimum. Wave physics analysis provide useful information to better assess what are the requirements that an initial model should satisfy to ensure the convergence toward the global minimum.

The non-convexity of the misfit function with respect to the P-wave velocity is related to the choice of the function $g(d_{cal}, d_{obs})$ to measure the discrepancy between observed and calculated data. Seismic observations are in essence oscillatory signals. Macro-scale P-wave velocity perturbation mainly affect the seismic data by modifying the propagation time rather than the amplitude of the seismic events [Jannane et al., 1989]. As a result, observed and calculated data mainly differ through time-shifts of the different seismic arrivals. The function $g(d_{cal}, d_{obs})$ should thus be convex with respect to these time-shifts. This is not the case for the L^2 distance which is used in practice. This is illustrated in Figure 1 where the seismic data is schematically represented as a periodic sinusoidal signal. When the signals are shifted by a multiple of one period of the signal, the L^2 differences between the signals reaches a local minimum: this is what is referred to a cycle skipping, or phase ambiguity problem, in the FWI community. Avoiding these local minima thus requires to start the minimization from less than half-a-phase shift. In other words, the initial velocity model should be sufficiently accurate to predict the kinematic of the wave propagation up to half-a-phase shift.

Mitigating this non-convexity has been the aim of numerous methods proposed during the past decades. Three main lines of investigation have been followed. The first one relies on the design of hierarchical schemes. The data is interpreted through a sequence of FWI problems, the estimation obtained from the problem i being used as an initial guess for the problem $i+1$. For each FWI problem, only a subset of the data is interpreted. The usual data decomposition is performed in the frequency domain: the data is interpreted from low-to-high frequencies. Low frequency components of the signal have a larger period, therefore the requirement on the initial model to fit the observed data within

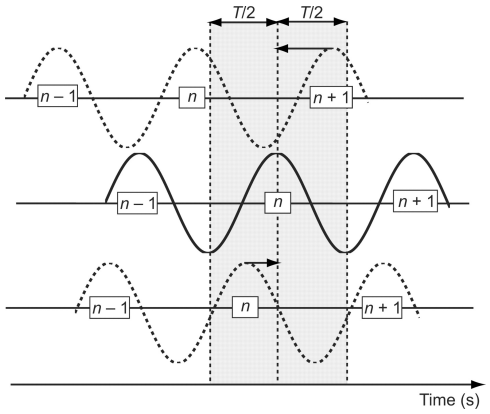

Fig. 1 Schematic example of the cycle skipping/phase ambiguity issue on sinusoidal signals. As soon as the initial shift is larger than half a period of the signal, the fit of the signal using a least-squares distance is performed up to one or several phase shifts. On may try to fit the $n+1$ dashed wriggle of the top signal with the n continuous wriggle of the middle signal moving to the wrong direction. The bottom dashed signal predicts the n wriggle in less than half-period leading to a correct updating direction (figure from [Virieux and Operto, 2009]).

half-a-period of the signal is partially relaxed. Additional level of hierarchy can also be applied (time-windowing, offset selection for instance) following layer stripping approaches [Bunks et al., 1995, Pratt, 1999, Shipp and Singh, 2002]. The second line of investigation is based on the modification of the misfit measurement function $g(d_{cal}, d_{obs})$. Cross-correlation functions have been first investigated [Luo and Schuster, 1991], and later on warping techniques [Hale, 2013], deconvolution approaches [Luo and Sava, 2011, Warner and Guasch, 2014], as well as envelope and phase separation [Fichtner et al., 2008, Bozdağ et al., 2011]. The third line of investigation relies on probing the consistency of the velocity model by building reflectivity images using different subset of the data. The velocity is updated such that the different reflectivity images become similar (see [Symes, 2008] and references therein for a review). These methods are known as (extended) image-domain techniques.

None of these approaches has completely overcome the cycle skipping or phase ambiguity problem. Hierarchical approaches relax the constraint on the accuracy of the initial velocity model by working first at low frequencies however this strategy is limited by the lowest available frequency, which is most of the time not low enough to sufficiently constrain the model. The different modifications of the misfit function proposed so far also enables to start from an initial velocity model further away from the solution, however this is often at the expense of the resolution of the final estimation. Image-domain techniques also exhibit interesting properties in terms of convexity of the misfit function, however, the computation cost associated with the repeated computation of reflectivity images seems to preclude their use to large-scale data-sets, especially in 3D configuration.

In this study, we discuss how optimal transport distances could be used to define an alternative misfit function measurement g in the framework of FWI. In particular, these distances provide natural tools to go beyond the point-to-point comparison underlaid by the least-squares distance by performing global comparison. The field of optimal transport has been very active in the last years, as testified by the number of textbooks published on this topic recently [Villani, 2003, Villani, 2008, Ambrosio et al., 2008, Santambrogio, 2015]. Recent applications in image processing demonstrate the interest of optimal transport distance to compare images, notably for its ability to detect shifted patterns from one image to another [Lellmann et al., 2014]. We discuss what are the main difficulties when applying optimal transport distance for the comparison of seismic data. In particular, we show that the oscillatory nature of the seismic data requires to extend optimal transport to the comparison of signed measures, which is a non-trivial problem. We review three different propositions found in the literature relying on the decomposition of the data in its positive and negative part. We show how the two first options might not be adapted for full waveform inversion. We thus focus on the third possibility, and show how an efficient implementation can be obtained, as we have presented it in previous studies [Métivier et al., 2016c, Métivier et al., 2016a]. We present numerical results obtained on the 2D Marmousi case study, a benchmark in the seismic imaging community, which illustrate the interest of this approach.

In Section 2, we discuss the optimal transport problem formulation for positive measures, and present a state-of-the-art for its extension to the comparison of signed measures. In Section 3, we present the alternative strategy we have promoted in previous studies and its application to the 2D Marmousi case study. Conclusion and perspectives are given in Section 4.

2 Optimal transport for full waveform inversion

2.1 Basics on optimal transport

Optimal transport has its roots in the work of a French scientist named Gaspard Monge, in an attempt to devise the best strategy to move piles of sand to a building site. The aim was to minimize the volume of the sand to be displaced, as well as the distance on which it had to be displaced. In modern mathematics, an expression of this problem is the following. Consider two probability measures $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$ (μ would represent the initial configuration of sand and ν the targeted one). We consider the mapping $T(x)$ from X to Y such that

$$\begin{cases} X & \longrightarrow & Y \\ T : x & \longrightarrow & T(x), \end{cases} \tag{12}$$

The push forward distribution of μ through the mapping T is denoted by $T_{\#}\mu$, such that for any measurable subset $A \subset Y$, we have

$$(T_{\#}\mu)(A) \equiv \mu\left(T^{-1}(A)\right) = \nu(A). \tag{13}$$

In this framework, the original Monge problem is formulated as

$$\inf_{T}\left\{\int_X ||x-T(x)||d\mu(x),\ \ T_{\#}\mu=\nu\right\}. \tag{14}$$

This problem has not necessarily a solution, and when the solution exists, it is difficult to compute because of the nonlinear constraint $T_{\#}\mu = \nu$.

A relaxation of this problem has been proposed by [Kantorovich, 1942], under the form

$$\inf_{\gamma}\left\{\int_{X\times Y} c(x,y)d\gamma(x,y),\ \ \gamma\in\Pi(\mu,\nu)\right\}, \tag{15}$$

where the ensemble of transport plans $\Pi(\mu,\nu)$ is defined by

$$\Pi(\mu,\nu) = \{\gamma\in\mathscr{P}(X\times Y),\ \ (\pi_X)_{\#}\gamma=\mu,\ \ (\pi_Y)_{\#}\gamma=\nu\}. \tag{16}$$

The operators π_X and π_Y are the projectors on X and Y respectively. This relaxation is the cornerstone of modern application of optimal transport as the problem (15) has always a solution which coincides with the one of the original Monge problem when this one exists. The problem (15) generalizes (14) in the sense that, instead of considering a mapping T transporting each particle of the distribution μ to the distribution ν, it considers all pairs (x,y) of the space $X\times Y$ and for each pair defines how many particles of μ go from x to y.

In discrete form, the Kantorovich problem becomes a linear programming problem of the form

$$\min_{\gamma_{ij}}\sum_{ij}\gamma_{ij}c_{ij},\ \ \gamma\in\Pi(\mu,\nu) \tag{17}$$

where

$$\Pi(\mu,\nu) = \{\gamma\geq 0, \sum_{j=1}\gamma_{ij}=\mu_i,\ \ \sum_{i=1}\gamma_{ij}=\nu_j\} \tag{18}$$

The entry γ_{ij} represents how much mass should be moved from x_i to y_j while c_{ij} measures the distance between x_i to y_j. The constraint ensures that the initial distribution is equal to μ while the transported distribution through the transport plan γ is equal to ν.

Of particular interest, optimal transport induces distances between distribution, named as Wasserstein distances or Earth's Mover Distances (EMD). They are defined by

$$W_p(\mu,\nu) = \left(\min_{\gamma_{ij}}\sum_{ij}\gamma_{ij}\|x_i-y_j\|^p,\ \ \gamma\in\Pi(\mu,\nu)\right)^{1/p} \tag{19}$$

One interest for using such distance for signal processing applications is their ability to detect shifted pattern from one signal to another. This property is also referred to in the literature as the fact that W_p distances should be seen as "horizontal distances" while L^p distances should be seen as "vertical distances" [Santambrogio, 2015]. The

W_p distance between two shifted probability distributions is convex with respect to this shift, while the L^p distance is insensitive to this shift.

2.2 Applying optimal transport for the comparison of seismic data: the difficulty of transporting signed measures

The existence of a solution to the optimal transport problem (16) depends on two assumptions that shall be satisfied by the measures μ and ν

1. μ and ν shall be positive
2. μ and ν shall have the same total mass

$$\int_X d\mu(x) = \int_X d\nu(x). \tag{20}$$

In this section, for the sake of simplicity, we assume that the two measures μ and ν are defined on the same space X . This is the case when μ and ν represent seismic data. Seismic data do not satisfy the positivity requirement due to its oscillatory nature. However, the zero frequency component of each seismic trace is zero

$$\forall x_r, \ \int_0^T d_{cal}(x_r,t)dt = \int_0^T d_{obs}(x_r,t)dt = 0. \tag{21}$$

Therefore we have

$$\int_{x_r}\int_0^T d_{cal}(x_r,t)dtdx_r = \int_{x_r}\int_0^T d_{obs}(x_r,t)dtdx_r = 0. \tag{22}$$

Thus, interpreting seismic data as density functions, equation (22) shows that the seismic data satisfy the second assumption: observed and calculated data have the same total mass, which is zero.

The main difficulty to apply optimal transport to the comparison of seismic data thus relies on the non-positivity of the seismic data. This is a well identified issue in the optimal transport community. The question how to extend optimal transport to signed measures is investigated in particular by [Ambrosio et al., 2008] and [Mainini, 2012]. Mainini makes use of the following Jordan-Hanh decomposition,

$$\mu = \mu^+ - \mu^-, \tag{23}$$

where μ^+ (respectively μ^-) is the positive part of μ (respectively the negative part of μ). Three strategies are reviewed in [Mainini, 2012] to extend optimal transport to signed measures. The corresponding extension of the W_p distances to signed measures are introduced as $W_{p,i}(\mu,\nu)$, $i = 1,2,3$ in the following. The three strategies proposed by Mainini are

1. Transport separately the positive and negative part of the measures

$$W_{p,1}(\mu,\nu) = W_p(\mu^+,\nu^+) + W_p(\mu^-,\nu^-) \tag{24}$$

2. Transport the absolute value of the measures

$$W_{p,2}(\mu,\nu) = W_p(|\mu|,|\nu|) \tag{25}$$

3. Perform the decomposition

$$W_{p,3}(\mu,\nu) = W_p(\mu^+ + \nu^-,\nu^+ + \mu^-) \tag{26}$$

The first strategy, which might appear as the more intuitive, is the one proposed originally by [Engquist and Froese, 2014]. It is successfully applied to the comparison of two time-shifted Ricker functions. The function $W_{2,1}^2(\mu,\nu)$ exhibits a quadratic convexity with respect to the time-shift between the two Ricker functions (Fig. 2). Two drawbacks can nonetheless be identified. First, the mass conservation between positive and negative parts of the measure is not ensured. Second,there is no obvious reason that the positive and negative parts of the seismic data should be uncorrelated. For realistic application, the source wavelet $s(x,t)$ is not known, and a prior source estimation is required to perform FWI. Hence, we can expect this decomposition to be strongly sensitive to errors in this source wavelet estimation.

The second strategy is straightforward to apply, however, the mass conservation between $|\mu|$ and $|\nu|$ is also not ensured. In addition, FWI misfit functions relying on the absolute value of the data lose the sensitivity to the polarity of the signal. As a result, positive or negative impedance contrasts can not be distinguished. This prevents from the correct interpretation of reflected waves.

The third strategy comes from the decomposition

$$\mu - \nu = \left(\mu^+ + \nu^-\right) - \left(\nu^+ + \mu^-\right). \tag{27}$$

This method seems appealing as, for any μ and ν satisfying the mass conservation assumption one has

$$\int_X d\mu^+ - d\mu^-(x) = \int_X d\nu^+(x) - d\nu^-(x), \tag{28}$$

therefore

$$\int_X d\mu^+ + d\nu^-(x) = \int_X d\nu^+(x) + d\mu^-(x), \tag{29}$$

and the mass conservation is ensured for the distance $W_{p,3}$.

We thus see that the mass conservation assumption is not satisfied in the definition of $W_{p,1}, W_{p,2}$. This might not be a shortcoming as severe as the one associated with the transport of signed measures as several possibilities exist to extend optimal transport to situation where the mass conservation is not ensured, known as partial optimal transport. However, the correlation between the negative and positive part of the seismic data is not accounted for using $W_{p,1}$. The sensitivity to the polarity of the seismic data is lost using $W_{p,2}$. These two drawbacks are severe. On the other hand, $W_{p,3}$ is based on a formulation for which the mass conservation is ensured and only

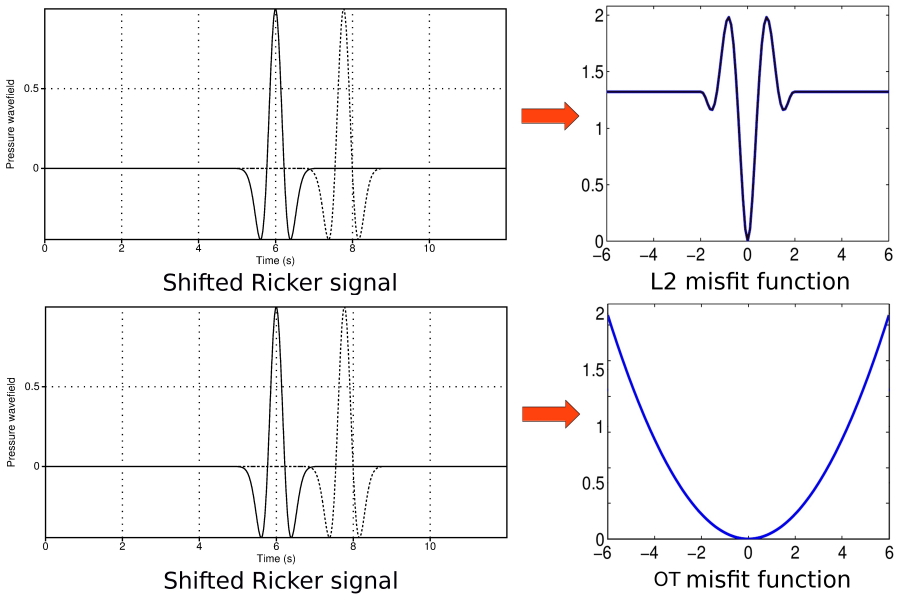

Fig. 2 Computation of the misfit function between two time-shifted Ricker signal depending on the time shift, using a least-squares distance and an optimal transport distance. While the least-squares distance yields a non-convex misfit function with two local minima aside the global minimum at zero time-shift, the optimal transport distance yields a perfectly convex misfit function [Engquist and Froese, 2014].

positive measures are compared. For this reason we are interested in investigating the use of this strategy for FWI.

2.3 A strategy using the W_1 distance in its dual form

2.3.1 Link between the dual W_1 distance and the Mainini decomposition

As the size of seismic data easily reaches several millions of discrete parameters for realistic FWI applications, we need to design a numerical strategy for large scale optimal transport problem with at most quasi-linear complexity.

Standard approaches for fast optimal transport computation encompass

- the direct solution of the Monge-Ampère equations [Philippis and Figalli, 2014]
- the solution of a fluid dynamic problem following the Benamou-Brenier formulation [Benamou and Brenier, 2000]
- the solution of a regularized optimal transport problem following an entropic regularization strategy [Cuturi, 2013, Benamou et al., 2015]

The last of this strategy can be applied for the computation of general W_p distances, while the two first strategies are dedicated to the computation of the W_2 distance.

Instead of relying on these developments, we rather propose another fast optimal transport computation technique, dedicated to a particular instance of the W_1 distance. The reason we focus on the W_1 distance is related to the Mainini technique, described in the previous paragraph, we want to apply. We explain it in the following.

The very important duality result due to [Kantorovich, 1942] states that the Kantorovich optimal transport problem (16) is equivalent to the maximization problem

$$\max_{\varphi,\psi} \int_X \varphi(x) d\mu(x) + \int_X \psi(x) d\nu(x), \;\; \varphi(x) + \psi(y) \leq c(x,y). \tag{30}$$

In the particular case of the W_1 distance, the dual problem (30) can be expressed using a single potential function $\varphi(x)$ as

$$\max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x) d(\mu - \nu)(x), \tag{31}$$

where the space of 1-Lipschitz function over X is denoted by $\mathrm{Lip}_1(X)$. This simplification comes from the fact that for W_1, we have

$$c(x,y) = |x-y| \tag{32}$$

which is itself a distance over $X \times X$ (see [Santambrogio, 2015] for a complete proof). Note that this is not the case for W_p distances with $p > 1$.

Interestingly, using this duality result, we see that

$$\begin{aligned} W_{1,3}(\mu,\nu) &= W_1(\mu^+ + \nu^-, \nu^+ + \mu^-) \\ &= \max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x) d(\mu^+ + \nu^- - \nu^+ + \mu^-)(x), \\ &= \max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x) d(\mu - \nu)(x) \\ &= W_1(\mu,\nu) \end{aligned} \tag{33}$$

This equality is important, as it reveals that through its particular dual formulation, the distance W_1 (31) can be computed for signed measures satisfying the mass conservation assumption (22). Indeed, as it is mentioned in [Lellmann et al., 2014] and [Bogachev, 2007][8.10.viii], the problem

$$\max_{\varphi \in Lip(X)} \int_X \varphi_x d\mu(x), \tag{34}$$

defines the norm $\|\mu\|^*_{KR}$ on the space of signed measures with first-order moment equal to zero

$$\int_X d\mu(x) = 0. \tag{35}$$

We have mentioned that for seismic data, the measure $\mu - \nu$ satisfies (35), therefore we have

$$\left\{ \max_{\varphi \in Lip_1(X)} \int_X \varphi(x) d(\mu - \nu)(x), \right\} = \|\mu - \nu\|_{KR}^* \tag{36}$$

In addition, this shows that the Mainini decomposition is directly embedded in the dual formulation of W_1 as soon as signed measures are involved.

This has the following important advantage for our application: there is no need to numerically perform the Jordan-Han decomposition into positive and negative part of the data to compute our misfit function. This could be problematic as we minimize this misfit function through local optimization strategies for differentiable functions, relying on the gradient and the Hessian of this function. As the Jordan-Han decomposition is not differentiable (by definition), the resulting misfit function would not be differentiable, and we would need to use optimization strategies for non-smooth misfit functions.

Note that in the case the mass conservation assumption is not satisfied, the norm $\|.\|_{KR}^*$ can be easily extended to the Kantorovich-Rubinstein norm, defined by

$$\|\mu - \nu\|_{KR} = \left\{ \max_{\varphi} \int_X \varphi(x) d(\mu - \nu)(x), \;\; \varphi(x) \in Lip_1(X), \|\varphi\|_\infty < 1 \right\} \tag{37}$$

This problem admits a solution even in the case $\mu - \nu$ does not satisfy (35). It might be more flexible to use for realistic application as the mass conservation is satisfied only at machine precision, which might occur instabilities when using the formulation (31).

In a series of articles [Métivier et al., 2016b, Métivier et al., 2016c, Métivier et al., 2016a], we have investigated the use of this Kantorovich-Rubinstein norm for realistic FWI applications. In the following, we summarize the numerical method developed in these studies to compute this norm.

2.3.2 Numerical method

We consider in the following the computation of the Kantorovich-Rubinstein norm for $d_{obs}(x_r,t) - d_{cal}(x_r,t)$. In discrete form, this is equivalent to the solution of the problem

$$\begin{aligned} &\max_{\varphi_{rn}} \sum_{r=1}^{N_r} \sum_{n=1}^{N_t} \varphi_{rn} \left((d_{obs})_{rn} - (d_{cal})_{rn} \right), \\ &\forall r,n,r',n' \;\; |\varphi_{rn} - \varphi_{r'n'}| \leq \|(x_r,t_n) - (x_r',t_n')\|, \\ &\forall r,n, \;\; |\varphi_{rn}| \leq 1 \end{aligned} \tag{38}$$

where the integer r is the index associated with the receiver coordinate x_r and the integer n is the index associated with the time coordinate t.

We denote by $N = N_r \times N_t$ the total number of discrete samples associated with one dataset. In this framework, the computation of the Kantorovich-Rubinstein norm is a linear programming problem with $O(N)$ unknowns and $O(N^2)$ constraints. For realistic application, N easily reaches $O(10^6)$, already for 2D problems. It is therefore important to reduce the number of constraints of the problem to reach feasible complexity algorithms.

With this purpose, we focus on the particular case where, instead of the Euclidean distance $\|.\|$, we use the ℓ_1 distance we denote by $|.|$ to measure the distance between (x_r,t_n) and (x'_r,t'_n). In [Métivier et al., 2016c] we show that satisfying the N^2 constraints

$$\forall r,n,r',n' \;\; |\varphi_{rn} - \varphi_{r'n'}| \leq |(x_r,t_n) - (x'_r,t'_n)| = |x_r - x'_r| + |t_n - t'_n| \tag{39}$$

is equivalent to satisfy the $2N$ constraints

$$\forall r,n \;\; |\varphi_{rn} - \varphi_{r+1,n}| \leq |x_r - x_{r+1}| \;\; |\varphi_{rn} - \varphi_{r,n+1}| \leq |t_n - t_{n+1}| \tag{40}$$

This is simply due to the "Manhattan" property of the ℓ_1 norm. This yields the following ℓ_1 Kantorovich-Rubinstein problem

$$\begin{aligned}
&\max_{\varphi_{rn}} \sum_{r=1}^{N_r} \sum_{n=1}^{N_t} \varphi_{rn} \left((d_{obs})_{rn} - (d_{cal})_{rn} \right), \;\; \forall r,n \\
&|\varphi_{rn} - \varphi_{r+1,n}| \leq |x_r - x_{r+1}| \\
&|\varphi_{rn} - \varphi_{r,n+1}| \leq |t_n - t_{n+1}| \\
&|\varphi_{rn}| \leq 1
\end{aligned} \tag{41}$$

which is a linear programming problem with $O(N)$ unknowns and $O(N)$ constraints.

In [Métivier et al., 2016c], we have detailed how this problem can be recast as the convex non-smooth optimization problem

$$\max_{\varphi} f_1(\varphi) + f_2(A\varphi), \tag{42}$$

where

$$f_1(\varphi) = \sum_{r=1}^{N_r} \sum_{n=1}^{N_t} \varphi_{rn} \left((d_{obs})_{rn} - (d_{cal})_{rn} \right), \;\; f_2(\psi) = i_K(\psi). \tag{43}$$

The function i_K is the indicator function on the unit hypercube K such that

$$i_K(x) = \begin{vmatrix} 0 & \text{if} \;\; x \in K \\ +\infty & \text{if} \;\; x \notin K, \end{vmatrix} \tag{44}$$

The operator A is the rectangular real matrix

$$A = \left[D_{x_r} \;\; D_t \;\; I_N \right]^T, \tag{45}$$

where I_N is the real identity matrix of size N and D_x, D_y, D_z are the forward finite differences operators

$$\begin{cases} (D_{x_r}\varphi)_{rn} = \dfrac{\varphi_{r+1,n} - \varphi_{rn}}{\Delta x_r}, \\ (D_t\varphi)_{rn} = \dfrac{\varphi_{r,n+1} - \varphi_{rn}}{\Delta t}, \end{cases} \tag{46}$$

Efficient strategies based on proximal splitting can be used to solve problems such as (42), where the functions f_i might not be differentiable. Among several possibilities, we choose the Simultaneous Direction Method of Multipliers (SDMM), which is well described in [Combettes and Pesquet, 2011], for its good convergence properties. The method can be summarized as the algorithm 1. The proximity operator

$\gamma > 0, y_1^0 = 0, y_2^0 = 0, z_1^0 = 0, z_2^0 = 0;$
for $n = 0, 1, \ldots$ **do**
  $\varphi^k = \left(I_N + A^T A\right)^{-1} \left[\left(y_1^k - z_1^k\right) + A^T \left(y_2^k - z_2^k\right)\right];$
  $y_1^{k+1} = \text{prox}_{\gamma f_1}\left(\varphi^k + z_1^k\right);$
  $z_1^{k+1} = z_1^k + \varphi^k - y_1^{k+1};$
  $y_2^{k+1} = \text{prox}_{\gamma_{i_K}}\left(A\varphi^k + z_2^k\right);$
  $z_2^{k+1} = z_2^k + A\varphi^k - y_2^{k+1};$
end

Algorithm 1: SDMM method for the solution of the problem (42).

can be seen as the generalization of the convex projection operator. For a given function f, it is defined as

$$\text{prox}_f(x) = \arg\min_y f(y) + \frac{1}{2}\|x - y\|_2^2, \tag{47}$$

For the particular case of the function f_1 and f_2, closed-form formulations exist

$$\text{prox}_{\gamma f_1}(\varphi) = \varphi - \gamma(d_{obs} - d_{cal}), \tag{48}$$

$$\forall i = 1, \ldots, P, \quad \left(\text{prox}_{\gamma f_2}(x)\right)_i = \left(\text{prox}_{i_K}(x)\right)_i = \begin{cases} x_i & \text{if } -1 \leq x_i \leq 1 \\ 1 & \text{if } x_i > 1 \\ -1 & \text{if } x_i < -1. \end{cases} \tag{49}$$

The closed-form formulations (48) and (49) are inexpensive to compute with an overall complexity in $O(N)$ operations. However, the SDMM algorithm requires the solution of a linear system involving the matrix $I + A^T A$. In [Métivier et al., 2016c], we show that the matrix $A^T A$ is a second-order finite-difference discretization of the Laplacian operator with homogeneous Neumann boundary conditions. Therefore, these linear systems can be solved in $O(NlogN)$ complexity using Fast Fourier Transform based algorithms [Swarztrauber, 1974], or in $O(N)$ complexity using multigrid strategies [Brandt, 1977, Adams, 1989]. The combination of the reduction of the number of constraints using the property of the ℓ_1 distance and the observa-

tion that the matrix $I + A^T A$ appearing in the SDMM strategy actually corresponds to the discretization of the Poisson's equation offers the possibility to design an efficient numerical method to compute the ℓ_1 Kantorovich-Rubinstein norm for large scale problems.

Following the notations used in Section 1, the use of the ℓ_1 Kantorovich-Rubinstein as the misfit measurement function for FWI implies that

$$g(d_{obs}, d_{cal}) = \|d_{cal} - d_{obs}\|_{KR} \tag{50}$$

The computation of the gradient of the resulting misfit function only requires the definition of the source of the adjoint field $\lambda_1(x,t)$ through

$$\frac{\partial \|d_{cal} - d_{obs}\|_{KR}}{\partial d_{cal}} \tag{51}$$

Interestingly, following the definition of $\|d_{cal} - d_{obs}\|_{KR}$, if we denote by $\overline{\varphi}$ the solution of the maximization problem (42), we have

$$\frac{\partial \|d_{cal} - d_{obs}\|_{KR}}{\partial d_{cal}} = \overline{\varphi} \tag{52}$$

As a consequence, the computation of the solution to the problem (42) yields simultaneously the value of the misfit function, through the value of the criterion at the maximum, as well as the quantity $\overline{\varphi}$ required to compute the gradient of the misfit function through the adjoint-state approach. The solution of a single optimal transport problem per seismic source is thus required at each iteration of FWI.

3 Example of application of the Kantorovich-Rubinstein norm to FWI

In order to illustrate the property of the Kantorovich-Rubinstein norm for the interpretation of seismic data, we first reproduce the experiment proposed in [Engquist and Froese, 2014] where the distance between shifted in time Ricker signal is computed using the L^2 distance and the W_2 distance applied to the positive and negative part of the Ricker separately. Here, instead of the W_2 distance, we compute directly the Kantorovich-Rubinstein distance without separating positive and negative parts of the signal. The results is presented in Figure 3. Compared to the least-squares distance, a single minimum is recovered. However, the convexity of the misfit function with respect to the time shift is lost. The loss of convexity is due to the signed nature of the Ricker signal (presence of negative values). One could expect optimal transport to be able to detect that the same pattern is shifted when comparing the Ricker, and that the W_1 distance would be proportional to this shift. This is not the case, which results from the presence of negative values. However, an important feature is preserved, with respect to the L^2 distance: a single minimum is obtained, while the L^2 distance

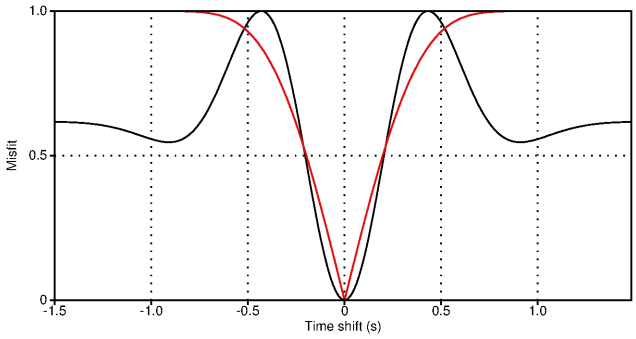

Fig. 3 Computation of the misfit function between two time-shifted Ricker signal depending on the time shift, using a least-squares distance (black) and the Kantorovich-Rubinstein distance (red). We recover a single minimum, however compared to the optimal transport distance used by [Engquist and Froese, 2014], the convexity of the misfit function is lost.

display two local minima aside the global minimum. This prompts us to test the use of the Kantorovich-Rubinstein norm to a more realistic case study.

To this purpose, we consider the Marmousi model presented in Figure 4(a). A synthetic dataset is computed in the 2D acoustic constant-density approximation. A fixed-spread surface acquisition is used, with 128 sources each 125 m and 168 receivers each 100 m, at 50 m depth. A Ricker source function centered on 5 Hz is used to generate the synthetic dataset. The frequency content of the source is high-pass filtered below 3 Hz to mimic realistic seismic data. In practical application, this frequency band is contaminated by noise, and therefore filtered out before inversion. Two initial P-wave velocity models are considered: the first contains the main features of the exact model, only with smoother interfaces (Fig. 4.b). The second is a strongly smoothed version of the exact model with very weak lateral variations and underestimated growth of the velocity in depth (Fig. 4.c). Starting from these two initial models, we compare the FWI results obtained using a least-squares distance and the ℓ_1 Kantorovich-Rubinstein distance. The minimization is performed using the l-BFGS algorithm [Nocedal, 1980] implemented in the SEISCOPE optimization toolbox [Métivier and Brossier, 2016].

These results are presented in Figure 4(d-g). Starting from the first initial model, a correct estimation of the P-wave velocity model is obtained, using both the L^2 distance (Fig. 4.d) and the ℓ_1 Kantorovich-Rubinstein distance (Fig. 4.f). The estimation of the low velocity zone at $x = 11$ km, $z = 2.5$ km is slightly improved using the ℓ_1 Kantorovich-Rubinstein distance, as a high velocity artifact located in this zone is computed using the the L^2 estimation. Starting from the second initial model, only the results obtained using ℓ_1 Kantorovich-Rubinstein distance are meaningful (Fig.4g). The poor initial approximation of the P-wave velocity is re-

sponsible for the cycle skipping effect and the L^2 estimation corresponds to a local minimum of the misfit function (Fig.4f). The estimation obtained with the ℓ_1 Kantorovich-Rubinstein distance is significantly closer from the true model, despite low velocity artifacts in the shallow part at $x = 1.5$ km, $z = 1$ km and in depth at $x = 12$ km, $z = 3.4$ km. This example illustrates the potential of optimal transport for

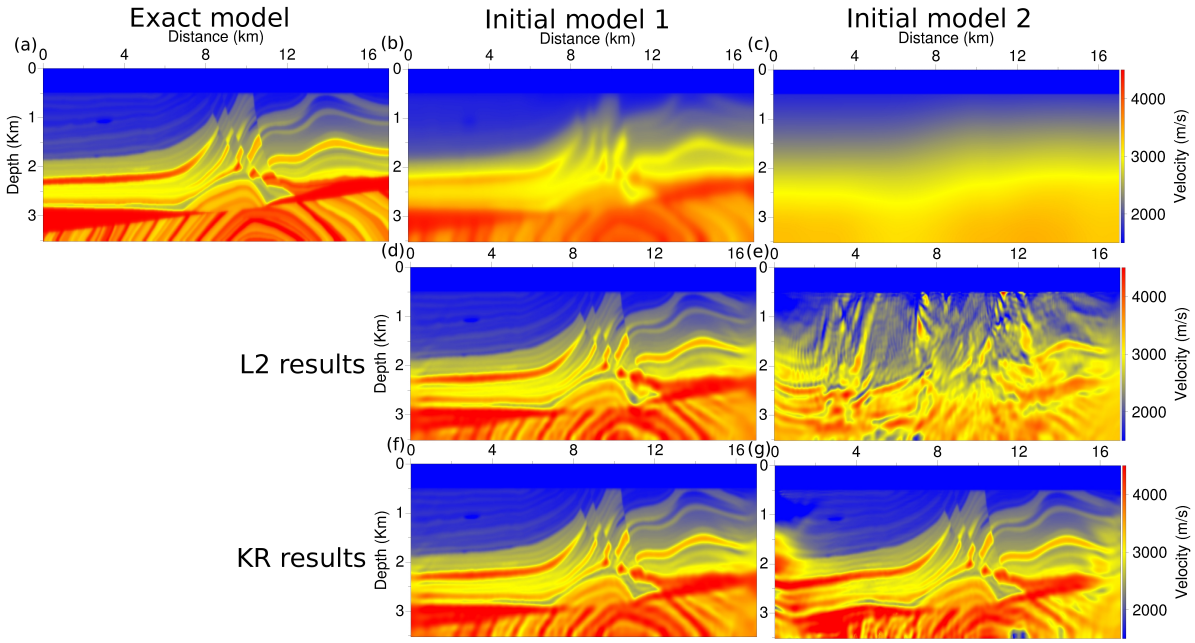

Fig. 4 Marmousi model case study. Exact model (a), initial model 1 (b), initial model 2 (c), results obtained with the L^2 distance starting from model 1 (d), from model 2 (e), results obtained with the ℓ_1 Kantorovich-Rubinstein distance starting from model 1 (f), from model 2 (g).

FWI: starting from a very crude approximation of the P-wave velocity, a meaningful estimation is computed. In the same configuration, FWI based on the least-squares distance fails and produces a heavily cycle skipped estimation.

4 Conclusion and Perspectives

The use of optimal transport distances for seismic imaging is promising. Comparing seismic data through these distances should yield more convex misfit functions with respect to the P-wave velocity parameter. However, the application of optimal transport to the comparison of seismic data requires the extension of the standard optimal transport problem to the transport of signed measures, which is not straightforward.

Standard decomposition techniques proposed in [Mainini, 2012], which are based on the negative and the positive part of the data, are either not adapted to FWI (separate transport of the positive and negative part, transport of the absolute value of the data), or lose the convexity property with respect to time-shifts which is one of the key properties one would like to satisfy for FWI.

Nonetheless, in the particular case of the dual formulation of the W_1 distance, the optimal transport distance can be related to a norm in the space of signed measure, the Kantorovich-Rubinstein norm. Hence, it can be directly use to compare seismic data. This is the strategy we have followed in previous works and which is summarized in this study. The results are encouraging. The resulting misfit function is not convex with respect to time-shifts, however, it already allows to start the FWI process from more crude initial velocity model, which denotes a wider valley of attraction of the misfit function. This method has been already successfully applied to 2D synthetic datasets in the context of deep water salt structures imaging (BP 2004 case study) and reflection dominated data (Chevron 2014 case study) [Métivier et al., 2016b], as well as to a 3D synthetic dataset (SEG/EAGE overthrust model) [Métivier et al., 2016c]. The method should now be applied to 2D and 3D real data-sets to further investigate the interest of this strategy for FWI.

Despite the interesting results provided by the Kantorovich-Rubinstein norm, the convexity property of the optimal transport distance with respect to shifted patterns on the data one could expect is lost. Further investigations are thus required to assess the feasibility of the design of a misfit function, based on optimal transport, adapted to the comparison of seismic data, which would benefit from this convexity property. Among different possibilities, one could think of the construction of positive observable from the seismic data, such as its envelope, which could thus be compared through W_p distances.

Acknowledgements This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by CGG, CHEVRON, EXXON-MOBIL, JGI, SHELL, SINOPEC, STATOIL, TOTAL and WOODSIDE. This study was granted access to the HPC resources of the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS/TGCC under the allocation 046091 made by GENCI.

References

[Adams, 1989] Adams, J. C. (1989). MUDPACK: Multigrid portable FORTRAN software for the efficient solution of linear elliptic partial differential equations. *Applied Mathematics and Computation*, 34(2):113–146.

[Ambrosio et al., 2008] Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

[Benamou and Brenier, 2000] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*.

[Benamou et al., 2015] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

[Bogachev, 2007] Bogachev, V. I. (2007). *Measure Theory*. Number vol. I,II in Measure Theory. Springer Berlin Heidelberg.

[Bozdağ et al., 2011] Bozdağ, E., Trampert, J., and Tromp, J. (2011). Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870.

[Brandt, 1977] Brandt, A. (1977). Multi-level adaptive solutions to boundary-value problems. *Mathematics of Computation*, 31:333–390.

[Bunks et al., 1995] Bunks, C., Salek, F. M., Zaleski, S., and Chavent, G. (1995). Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473.

[Chavent, 1971] Chavent, G. (1971). *Analyse fonctionnelle et identification de coefficients répartis dans les équations aux dérivées partielles*. PhD thesis, Université de Paris.

[Combettes and Pesquet, 2011] Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., and Wolkowicz, H., editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optimization and Its Applications*, pages 185–212. Springer New York.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*.

[Devaney, 1984] Devaney, A. (1984). Geophysical diffraction tomography. *Geoscience and Remote Sensing, IEEE Transactions on*, GE-22(1):3–13.

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

[Fichtner et al., 2008] Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H. P. (2008). Theoretical background for continental- and global-scale full-waveform inversion in the time-frequency domain. *Geophysical Journal International*, 175:665–685.

[Hale, 2013] Hale, D. (2013). Dynamic warping of seismic images. *Geophysics*, 78(2):S105–S115.

[Jannane et al., 1989] Jannane, M., Beydoun, W., Crase, E., Cao, D., Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Roeth, G., Singh, S., Snieder, R., Tarantola, A., and Trezeguet, D. (1989). Wavelengths of Earth structures that can be resolved from seismic reflection data. *Geophysics*, 54(7):906–910.

[Kantorovich, 1942] Kantorovich, L. (1942). On the transfer of masses. *Dokl. Acad. Nauk. USSR*, 37:7–8.

[Lailly, 1983] Lailly, P. (1983). The seismic inverse problem as a sequence of before stack migrations. In Bednar, R. and Weglein, editors, *Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia*, pages 206–220.

[Le Dimet and Talagrand, 1986] Le Dimet, F. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110.

[Lellmann et al., 2014] Lellmann, J., Lorenz, D., Schönlieb, C., and Valkonen, T. (2014). Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859.

[Lions, 1968] Lions, J. L. (1968). *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Dunod, Paris.

[Luo and Sava, 2011] Luo, S. and Sava, P. (2011). A deconvolution-based objective function for wave-equation inversion. *SEG Technical Program Expanded Abstracts*, 30(1):2788–2792.

[Luo and Schuster, 1991] Luo, Y. and Schuster, G. T. (1991). Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653.

[Mainini, 2012] Mainini, E. (2012). A description of transport cost for signed measures. *Journal of Mathematical Sciences*, 181(6):837–855.

[Métivier and Brossier, 2016] Métivier, L. and Brossier, R. (2016). The SEISCOPE optimization toolbox: A large-scale nonlinear optimization library based on reverse communication. *Geophysics*, 81(2):F11–F25.

[Métivier et al., 2016a] Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016a). Increasing the robustness and applicability of full waveform inversion: an optimal transport distance strategy. *The Leading Edge*, 35(12):1060–1067.

[Métivier et al., 2016b] Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016b). Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Journal International*, 205:345–377.

[Métivier et al., 2016c] Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016c). An optimal transport approach for seismic tomography: Application to 3D full waveform inversion. *Inverse Problems*, 32(11):115008.

[Nash, 2000] Nash, S. G. (2000). A survey of truncated Newton methods. *Journal of Computational and Applied Mathematics*, 124:45–59.

[Nocedal, 1980] Nocedal, J. (1980). Updating Quasi-Newton Matrices With Limited Storage. *Mathematics of Computation*, 35(151):773–782.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.

[Operto et al., 2013] Operto, S., Brossier, R., Gholami, Y., Métivier, L., Prieux, V., Ribodetti, A., and Virieux, J. (2013). A guided tour of multiparameter full waveform inversion for multicomponent data: from theory to practice. *The Leading Edge*, Special section Full Waveform Inversion(September):1040–1054.

[Philippis and Figalli, 2014] Philippis, G. D. and Figalli, A. (2014). The monge-ampre equation and its link to optimal transportation. *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY*.

[Plessix, 2006] 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(2):495–503.

[Pratt, 1999] Pratt, R. G. (1999). Seismic waveform inversion in the frequency domain, part I : theory and verification in a physical scale model. *Geophysics*, 64:888–901.

[Santambrogio, 2015] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.

[Shipp and Singh, 2002] Shipp, R. M. and Singh, S. C. (2002). Two-dimensional full wavefield inversion of wide-aperture marine seismic streamer data. *Geophysical Journal International*, 151:325–344.

[Swarztrauber, 1974] Swarztrauber, P. N. (1974). A Direct Method for the Discrete Solution of Separable Elliptic Equations. *SIAM Journal on Numerical Analysis*, 11(6):1136–1150.

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

[Tarantola, 1984] Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266.

[Villani, 2003] Villani, C. (2003). *Topics in optimal transportation*. Graduate Studies In Mathematics, Vol. 50, AMS.

[Villani, 2008] Villani, C. (2008). *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin.

[Virieux et al., 2017] Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A., and Zhou, W. (2017). An introduction to Full Waveform Inversion. In Grechka, V. and Wapenaar, K., editors, *Encyclopedia of Exploration Geophysics*, chapter in press, page accepted. Society of Exploration Geophysics.

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

[Warner and Guasch, 2014] Warner, M. and Guasch, L. (2014). Adaptative waveform inversion - FWI without cycle skipping - theory. In *76th EAGE Conference and Exhibition 2014*, page We E106 13.