A graph space optimal transport distance as a generalization of Lp distances: application to a seismic imaging inverse problem

Ludovic Métivier, Romain Brossier, Quentin Merigot, Edouard Oudet

A graph space optimal transport distance as a generalization of L^p distances: application to a seismic imaging inverse problem

L. Métivier[1,2], R. Brossier[2], Q. Mérigot[3], E. Oudet[1]

[1] Laboratoire Jean Kuntzmann (LJK), Univ. Grenoble Alpes, CNRS, France
[2] Institut des sciences de la Terre (ISTerre), Univ. Grenoble Alpes, France
[3] Laboratoire de mathmatiques d'Orsay (LMO), Univ. Paris-Sud, France

E-mail: ludovic.metivier@univ-grenoble-alpes.fr

February 2019

Abstract. Optimal transport distance is an appealing tool to measure the discrepancy between datasets in the frame of inverse problems, for its ability to perform global comparisons and its convexity with respect to shifted patterns in the compared quantities. However, solving inverse problems might require to compare signed quantities, while the optimal transport theory has been developed for the comparison of probability measures. In this study we propose to circumvent this difficulty by applying optimal transport to the comparison of the graph of the data rather than the data itself. We investigate this approach in the frame of seismic imaging, where each channel of the oscillatory data is interpreted as a discrete point cloud. We demonstrate that the corresponding misfit function can be computed through the solution of series of linear sum assignment problem (LSAP), while, based on the adjoint state technique, its gradient can be computed from the assignment solution of these LSAP. We illustrate how this approach yields a convex misfit function in the frame of seismic imaging using the full waveform. We show how an efficient strategy, based on a specific LSAP solver, the auction algorithm, can be designed. We illustrate the interest of the approach on a realistic 2D visco-acoustic seismic imaging problem. The proposed strategy relaxes the constraint on the accuracy of the initial model, outperforming the conventional least-squares approach and a previously proposed optimal transport based approach. The computational time increases by only a few percents compared with the least-squares approach, opening the way to applications to 3D field data in the near future.

Keywords: seismic imaging, full waveform inversion, optimal transport, linear assignment problem, auction algorithm

Submitted to: *Inverse Problems*

1. Introduction

The framework of this study is a seismic imaging inverse problem named Full waveform inversion (FWI). It is based on the iterative minimization of a misfit function between observed and calculated data, over a space of model parameters which describe the subsurface. This method is used at various scales: from global to regional scales in seismology (Fichtner et al., 2010; Tape et al., 2010; Bozdağ et al., 2016), for seismic exploration in the oil & gas industry (Plessix and Perkins, 2010; Stopin et al., 2014; Operto et al., 2015), and at the near surface scale for geotechnical targets (Bretaudeau et al., 2013; Groos et al., 2014; Schäfer et al., 2013). The main interest of FWI compared to standard tomography methods is its high-resolution power, down to half the shortest wavelength of the propagated signals (Devaney, 1984). An up-to-date review of FWI and its applications can be found in Virieux et al. (2017).

One of the main limitations of FWI is related to the non-convexity of the least-squares misfit function which is conventionally used to define the distance between observed and calculated data. This non-convexity is a severe limitation because the minimization process is based on local optimization solvers: the expected resolution leads to use rather fine discretization of the model space, responsible for number of unknowns from $O(10^6)$ in 2D to $O(10^9)$ in 3D for realistic size FWI applications. For this reason, successful applications of FWI strongly rely on the definition of an accurate enough initial model, to ensure the convergence towards the global minimum of the misfit function.

The non-convexity of the least-squares misfit function has a physical interpretation. Large-scale to medium-scale perturbations of the subsurface velocities are mainly responsible for shifting in time the seismic data (Jannane et al., 1989). However, the least-squares misfit function is not convex with respect to time-shifts. Schematically, for two shifted sinusoidal signals of period T, minimizing the least-squares misfit between these two signals through local optimization will conduct to put them in phase only if the initial shift is smaller than $T/2$. As soon as the initial shift is larger than $T/2$, the minimization leads to interpreting the signals with at least one phase shift (Virieux and Operto, 2009). This phenomenon is referred to as phase ambiguity or cycle skipping. In terms of velocity model reconstruction, this phase shift can be interpreted as wrongly increasing or decreasing the velocity model to fit the data.

This difficulty has been reported since the introduction of FWI in the 80s (Gauthier et al., 1986). Overcoming this limitation has been the subject of an important number of studies. Introducing a hierarchy in the data interpretation is a standard way to proceed. The general aim is to include only few wavelengths of the data at each inversion step, to reduce the phase ambiguity. This is performed by focusing first on the low-frequency part of the data and/or selected events through time/offset windowing techniques (Bunks et al., 1995; Pratt, 1999; Shipp and Singh, 2002; Wang and Rao, 2009; Brossier et al., 2009). However, except for global and regional scale applications, the low-frequency part of the data is often not energetic enough for this strategy to be sufficient. In addition, designing this data hierarchy is strongly application-dependent and requires specific human expertise.

Another - possibly complementary - strategy consists in designing more convex misfit functions, presenting less local minima than the least-squares misfit function, with a wider valley of attraction near the global minimum. This can be done working directly in the data space, modifying the measure of distance between two data sets. Cross-correlation (Luo and Schuster, 1991; van Leeuwen and Mulder, 2010), deconvolution (Luo and Sava, 2011; Warner and Guasch, 2016), instantaneous phase or envelope (Fichtner et al., 2008; Bozdağ et al., 2011) have been proposed for instance. Other convexification strategies rely on extending the parameter space. In migration velocity analysis (MVA), a redundant parameter is introduced in the image condition (a time-lag or subsurface offset for instance) to build image hypercubes. The correct velocity is expected to concentrate the energy at zero offset/time-lag. A new misfit function can thus be defined to penalize the energy away from zero in this augmented space. This function should behave as a travel-time tomographic operator and therefore should exhibit a better convexity (Symes, 2008, 2015). Another class of

extended space approach is referred to as wavefield reconstruction inversion (WRI) (van Leeuwen and Herrmann, 2016). This method is based on a relaxation on the constraint imposed to the wavefield to satisfy the wave equation to machine precision. In WRI, the wavefield becomes a new unknown. The resulting problem can be solved in an alternate way between the subsurface parameters and the wavefield (Aghamiry et al., 2018). It appears that this makes it possible to mitigate the non-convexity of the least-squares misfit function.

While all these strategies provide interesting alternative to the standard FWI formulation, they still suffer from a number of limitations. Cross-correlation based methods do not handle complex data sets including multiple and mixed seismic arrivals. The deconvolution strategy requires the tuning of a penalization function which might be application dependent. Instantaneous envelope requires also to carefully adjust scaling parameters and suffer from losing the polarity information, making difficult to correctly interpret the reflected events in the data. MVA methods are computationally intensive as the construction of seismic image hypercube, involving high-resolution migration, is required at each iteration of the method. WRI has been introduced for 2D acoustic frequency-domain FWI, and its generalization to 3D time-domain and elastic modeling is still under investigation (Wang et al., 2016).

Recently, an alternative data domain strategy based on the measure of the misfit between observed and synthetic data through optimal transport (OT) has been introduced (Engquist and Froese, 2014). OT is a mathematical field originating from the work of the French mathematician Gaspard Monge (Monge, 1781). The problem posed was to minimize the efforts performed by workers to transfer sand piles to fill in holes on a bridge building site. The corresponding minimization problem formulated by Monge is not well posed: a solution does not always exist. A well-posed relaxation of the problem was later proposed by Kantorovich (1942), leading to the definition of the OT (Wasserstein) distance: The solution of the OT problem, through the Kantorovich relaxation, defines a distance in the space of probability distributions.

This distance has a very interesting property: it is convex with respect to shifted patterns in the distributions which are compared. This property has made OT a widely used tool in image processing for applications such as image retrieval (Rubner et al., 2000; Rabin et al., 2010), histogram equalization (Delon, 2006), color transfer (Pitié et al., 2007), texture mapping (Dominitz and Tannenbaum, 2010; Rabin et al., 2012). More references on image processing applications of OT can also be found in Lellmann et al. (2014). In the field of seismic imaging, this property is very attractive. Assuming OT could be applied directly to seismic data, its convexity with respect to shifted patterns would produce a distance convex with respect to time shifts, a good proxy for the convexity with respect to the subsurface velocities. More generally, it also makes possible a global comparison of the seismic data, besides the quantitative point-to-point comparison induced by the use of L^p distances.

However, the seismic data is oscillatory, and cannot be represented as a probability distribution. The generalization of OT to signed measures is a fundamental mathematics problem. Until now, there are no clear ideas on how this extension could be formally made in a generic way (Ambrosio et al., 2011; Mainini, 2012). The application of OT to seismic data therefore relies on specific adaptations. A first approach consists in transforming the data into positive quantities and normalizing it prior being compared through OT (Engquist and Froese, 2014; Qiu et al., 2017; Yang et al., 2018b; Yang and Engquist, 2018). A second approach, which we have promoted, consists in considering a special instance of the OT distance, based on the 1-Wasserstein distance, which can be extended naturally to the comparison of non-positive data (Métivier et al., 2016a,b,c). We have proposed a review of these two types of adaptations in Métivier et al. (2018). We illustrate that both these types of strategies are either not adapted to the application to field data, or lose the convexity property that first motivates the use of OT for FWI.

We provide in Métivier et al. (2018) a solution to overcome this contradiction. We propose to compare the discrete graph of the data rather than the data itself, following the idea of Thorpe et al. (2017). Each trace (time-signal) constituting the data is interpreted as a point cloud after

a time-discretization. Mathematically, it is represented as a sum of Dirac delta functions in a 2D space containing the time-dimension and an additional dimension associated with the amplitude of the data. The discrete graph of the observed and the synthetic data is compared instead of the data itself. This guarantees the positivity of the compared quantities, and maintain the convexity of the distance with respect to time shifts and amplitude shifts. The illustrations provided in Métivier et al. (2018) are encouraging: the convexity of the misfit function is enhanced compared with previous implementation of OT based on the 1-Wasserstein distance. However, the introduction of an augmented data space can increase significantly the computational cost of the method. In Métivier et al. (2018), the proposed numerical strategy solves a fully 2D OT problem for each seismic trace, using the algorithm developed in Métivier et al. (2016c), making it prohibitively expensive for field data applications. For a 2D synthetic example, the additional computational cost for one iteration represents 700% compared to the least-squares distance.

The main contribution of this study is to provide a mathematical analysis of the graph space OT strategy, which yields the possibility to design a very efficient numerical strategy for this approach. In the numerical example presented here, the additional computational cost represents a 6 % increase compared to a least-squares based FWI algorithm. Interestingly, we demonstrate how the graph space OT distance can be interpreted as a generalization of conventional L^p distances.

We show in Section 2 how the computation of the graph space OT distance can be expressed as a linear sum assignment problem (LSAP), for which exists specific numerical solvers. In Section 3, we introduce this formalism in the frame of full waveform inversion. The analogy between the graph space OT distance and the L^p distance is highlighted. We prove how the adjoint source corresponding to the graph space OT misfit function is a generalization of the L^p adjoint source. We discuss practical implementation issues regarding the scaling between time and amplitude axis when considering 2D shot gathers. We then illustrate the convexity of the resulting misfit function on canonical examples. We provide an analysis of the gradient building with this new metric and establish a connection with another misfit function modification based on the cross-correlation of synthetic and observed data. In Section 4, we provide a numerical strategy for the computation of the misfit function and its gradient which takes benefit of the LSAP formulation, based on the auction algorithm (Bertsekas and Castanon, 1989). We show how the computational cost should marginally increase compared to L^p misfit functions, for realistic size FWI problems, making the method suitable for applications to 2D and 3D field data. In Section 5, we illustrate the interest of the approach on a realistic 2D synthetic benchmark based on the Valhall model, representative of the North Sea geology. Conclusion and perspectives are provided in the final Section.

2. Optimal transport distance between point clouds: a linear sum assignment problem

In this section we recall the basic definition of the Wasserstein distance through the Kantorovich relaxation problem. We show how its application to the comparison of discrete point clouds translates into a LSAP. We refer the readers to Villani (2008); Ambrosio (2003); Santambrogio (2015) for a more detailed introduction on the OT theory.

2.1. Basics on optimal transport theory

Consider two probability measures $\mu(x) \in \mathcal{P}(X)$ and $\nu(y) \in \mathcal{P}(Y)$, where X and Y are two complete and separable metric spaces. For a given map $T : X \to Y$, the image measure of μ through T is denoted by $T_{\#}\mu$, such that for any measurable set $A \subset Y$

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

The general definition of the Kantorovich problem is

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{X\times Y} c(x,y) d\gamma(x,y), \tag{2}$$

where $\Pi(\mu,\nu)$ is

$$\Pi(\mu,\nu) = \left\{\gamma \in \mathcal{P}(X\times Y),\ \ (\pi_X)_{\#}\,\gamma = \mu,\ \ (\pi_Y)_{\#}\,\gamma = \nu\right\}, \tag{3}$$

with π_X and π_Y the projections on X and Y respectively, and $c(x,y)$ a continuous function from $X\times Y$ to $[0;+\infty]$. The coupling measure γ specifies for each pair of points (x,y) how much particles from $\mu(x)$ should be moved to $\nu(y)$. The constraints (3) on $(\pi_X)_{\#}\,\gamma$ and $(\pi_Y)_{\#}\,\gamma$ make sure that γ describes a mapping from $\mu(x)$ to $\nu(y)$. The criterion which is minimized in (2) measures, for a given mapping (or transport plan) $\gamma(x,y)$, the total energy required to achieve this mapping. It is the sum, for all points (x,y), of the product between the amount of mass $\gamma(x,y)$ which is moved, multiply by a measure of the distance between these points $c(x,y)$. Solving the Kantorovich problem amounts to find the optimal transport plan $\gamma(x,y)$ which minimizes this criterion.

The Kantorovich problem can be used to define a distance between probability measures, named the Wasserstein distance. Assuming that

$$X = Y = \Omega \subset \mathbb{R}^n,\ \ n \in \mathbb{N}, \tag{4}$$

and that μ and ν are probability measures with finite p-moment, $p \in \mathbb{N}$, such that

$$\int_\Omega \|x\|^p d\mu(x) < +\infty,\ \ \int_\Omega \|x\|^p d\nu(x) < +\infty, \tag{5}$$

with $\|.\|$ a norm on $\mathbb{R}^n$, the p-Wasserstein distance between μ and ν is defined as

$$W_p(\mu,\nu) = \left(\min_{\gamma\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} \|x - x'\|^p d\gamma(x,x')\right)^{1/p}. \tag{6}$$

In this study, we will consider the case where $\|.\|.$ is the Euclidean norm on $\mathbb{R}^n$.

2.2. Discrete Kantorovich problem and comparison of point clouds

In practice, we deal with discrete instances of the p-Wasserstein distance. Considering the domain Ω is discretized as

$$\Omega = \{x_i,\ i = 1,\dots,N\}, \tag{7}$$

we introduce the conventional notations

$$\gamma(x_i,x_j) = \gamma_{ij},\ \mu(x_i) = \mu_i,\ \nu(x_j) = \nu_j. \tag{8}$$

In these discrete settings, the p-Wasserstein distance between μ and ν can be computed through the solution of the linear programming problem

$$\begin{aligned} W_p(\mu,\nu)^p = \min_{\gamma_{ij}} \quad & \sum_{i,j=1}^{N} \gamma_{ij}\|x_i - x_j\|^p, \\ & \gamma_{ij} \geq 0, \qquad i,j = 1,\dots,N, \\ & \sum_{j=1}^{N} \gamma_{ij} = \mu_i, \quad i = 1,\dots,N, \\ & \sum_{i=1}^{N} \gamma_{ij} = \nu_j, \quad j = 1,\dots,N. \end{aligned} \tag{9}$$

For two point clouds, the two distributions $\mu(x)$ and $\nu(y)$ can be described as a normalized sum of Dirac delta functions $\delta(x)$, such that

$$\mu(x) = \frac{1}{K}\sum_{k=1}^{K} \delta(x - u_k), \;\; \nu(x) = \frac{1}{K}\sum_{k=1}^{K} \delta(x - v_k), \tag{10}$$

where $u_k \in \Omega$ (respectively $v_k \in \Omega$) indicates the position of the points defining $\mu(x)$ (respectively $\nu(x)$). Denote

$$I_K = \{i, \;\; \exists k \in \{1,\dots,K\}, \;\; i = i_k\}, \;\; J_K = \{j, \;\; \exists k \in \{1,\dots,K\}, \;\; j = j_k\}, \tag{11}$$

where i_k and j_k denote the position on the discrete mesh of the points u_k and v_k respectively. Consider

$$\overline{I_K} = \{1,\dots,N\}/I_K, \;\; \overline{J_K} = \{1,\dots,N\}/J_K. \tag{12}$$

We have

$$\forall i \in \overline{I_K}, \;\; \mu_i = 0, \;\; \forall j \in \overline{J_K}, \;\; \nu_j = 0. \tag{13}$$

Because of the positivity constraints on γ_{ij}, this implies that all the coefficients of the rows $i \in \overline{I_K}$ and the columns $j \in \overline{J_K}$ of γ are equal to 0. These coefficients thus have no contribution in the criterion and can be removed. The linear programming problem (9) can thus be simplified as

$$W_p(\mu,\nu)^p = \min_{\gamma_{ij}} \sum_{i,j=1}^{K} \gamma_{ij}\|u_i - v_j\|^p \tag{14a}$$

$$\gamma_{ij} \geq 0, \qquad i,j = 1,\dots,K, \tag{14b}$$

$$\sum_{i=1}^{K} \gamma_{ij} = \frac{1}{K}, \quad j = 1,\dots,K, \tag{14c}$$

$$\sum_{j=1}^{K} \gamma_{ij} = \frac{1}{K}, \quad i = 1,\dots,K. \tag{14d}$$

or equivalently

$$W_p(\mu,\nu)^p = \frac{1}{K}\min_{\gamma_{ij}} \sum_{i,j=1}^{K} \gamma_{ij}\|u_i - v_j\|^p \tag{15a}$$

$$\gamma_{ij} \geq 0, \qquad i,j = 1,\dots,K, \tag{15b}$$

$$\sum_{i=1}^{K} \gamma_{ij} = 1, \quad j = 1,\dots,K, \tag{15c}$$

$$\sum_{j=1}^{K} \gamma_{ij} = 1, \quad i = 1,\dots,K. \tag{15d}$$

The linear programming problem (15) is a standard linear sum assignment problem (LSAP). While, formally, fractional values for the unknown γ_{ij} are authorized, a standard result states that the values of the optimal solution are either 0 or 1. This comes from the fact that the matrix describing the linear constraints (15c) and (15d) is totally unimodular: for such a linear programming problem, the solution can take only integer values (Birkhoff, 1946; Burkard et al., 2012).

Using this result, the problem (15) can be further reformulated as

$$W_p(\mu,\nu)^p = \frac{1}{K} \min_{\sigma \in S(K)} \sum_{i=1}^{K} \|u_i - v_{\sigma(i)}\|^p, \tag{16}$$

where $S(K)$ denotes the ensemble of permutations of $\{1,\ldots,K\}$.

3. Full waveform inversion using optimal transport in the graph space

In this section, we first introduce the mathematical framework for FWI as a PDE-constrained optimization problem. In its conventional formulation, the data fitting term is based on the L^p distance. We then define a new FWI strategy based on a graph-space OT misfit function. We show how this strategy can be understood as a generalization of the conventional L^p distance based FWI, by highlighting the connections between the formulas defining the misfit functions and their gradients. We then discuss practical implementation issues related to the amplitude and time axis scaling, especially when considering the interpretation of 2D shot gathers. We then explore the properties of this FWI strategy through canonical examples implying the comparison of shifted Ricker signal, the computation of 2D misfit function maps, and the analysis of the residuals in a simple transmission case.

3.1. Conventional L^p based FWI

A conventional seismic acquisition comprises at set of seismic sources (pressure or directional forces) and a set of receivers (hydrophones or geophones), generally located at or near the free surface. For each source, the receivers record the wavefield during a certain time duration $T \in \mathbb{R}_+^*$. The resulting seismic dataset can be represented as a collection of time dependent functions named seismic traces $d^{s,r}(t)$ where $s \in \mathbb{N}$ and $r \in \mathbb{N}$ correspond respectively to the source and receiver indexes, and $t \in \mathbb{R}_+$ denotes the time variable.

After time discretization, a seismic trace is a vector of K time samples $[d_1^{s,r}, \dots, d_K^{s,r})] \in \mathbb{R}^K$ where we use the notation

$$d^{s,r}(t_k) = d_k^{s,r}, \quad k = 1, \dots, K. \tag{17}$$

FWI is based on the comparison between observed and synthetic data $d_{obs}^{s,r}$ and $d_{cal}^{s,r}$. The synthetic data is computed in two steps. First, the wavefield $u^s(x,t)$ corresponding to the source s is computed through the solution of partial differential equations (PDE) representing the wave propagation within the subsurface. Under a general formulation, this set of PDE can be written as

$$A(m, \partial_t, \partial_x)u(x,t) = f^s(x,t), \quad (x,t) \in \mathbb{R}^d \times [0,T], \tag{18}$$

where $A(m, \partial_t, \partial_x)$ is a time-space partial differential operator, $m(x)$ represents the subsurface parameters, d is the spatial dimension ($d = 1, 2, 3$). Initial homogeneous conditions are imposed, as the medium is considered at rest before the propagation. For the boundary conditions, we usually assume zero energy at infinity, which can be efficiently implemented through absorbing layers techniques (Cerjan et al., 1985; Bérenger, 1994; Métivier et al., 2014).

Based on the solution of this PDE, the calculated data $d_{cal}^{s,r}(t)$ is given by

$$d_{cal}^{s,r}(t) = u^s(x_r, t), \tag{19}$$

where $x_r \in \mathbb{R}^d$ denotes the spatial position of the r^{th} receiver.

After discretization, the extraction of the wavefield values at the receivers positions can be written as the linear operation

$$d_{cal}^{s,r} = R_r u^s, \tag{20}$$

where now $d_{cal}^{s,r} \in \mathbb{R}^K$, $u^s \in \mathbb{R}^{NK}$, with N the number of discrete samples in space, and $R_r \in \mathbb{M}_{K,NK}(\mathbb{N})$ is a sparse matrix defined as a discretization of the Dirac comb extracting for each time sample the value of the wavefield u at the receiver position x_r at time t_k.

For $(d_{cal}, d_{obs}) \in \mathbb{R}^K \times \mathbb{R}^K$, we introduce the L^p based comparison function

$$h_{L^p}(d_{cal}, d_{obs}) = \frac{1}{p}\sum_{i=1}^{K} |d_{cal,i} - d_{obs,i}|^p = \frac{1}{p}\|d_{cal} - d_{obs}\|_p^p, \tag{21}$$

where $\|.\|_p$ is the L^p norm in $\mathbb{R}^K$. The conventional L^p FWI problem is formulated as the minimization problem

$$\min_m f_{L^p}[m] = \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} h_{L^p}(d_{cal}^{s,r}[m], d_{obs}^{s,r}), \tag{22}$$

where the dependency of the calculated data with respect to the velocity model $m(x)$ is denoted with square brackets $[m]$. In practice, mostly the cases $p = 2$ (least-squares) or $p = 1$ are considered, the latter being preferred to interpret heavily noise-contaminated data for its better robustness with respect to outliers (see Brossier et al., 2010, for instance).

The minimization problem (22) is solved through quasi-Newton techniques (Nocedal, 1980) based on the iteration

$$m^{q+1} = m^q + \alpha^q \Delta m^q, \tag{23}$$

starting from a given m^0. In (23), $\alpha^q \in \mathbb{R}^+$ is a linesearch parameter ensuring the convergence towards the closest local minimum, and Δm^q is a descent direction, given by

$$\Delta m_q = -B_q \nabla f[m_q], \tag{24}$$

where B_q is the l-BFGS approximation of the inverse Hessian operator computed from l previous gradients $\nabla f[m_{q-l+1}], \ldots, \nabla f[m_q]$ (see Métivier and Brossier, 2016, for a review of local optimization techniques for FWI). The key quantity to compute is thus the gradient $\nabla f_{L^p}[m]$.

Following the adjoint-state approach, the gradient $\nabla f_{L^p}[m]$ can be expressed as

$$\nabla f_{L^p}[m] = \sum_{s=1}^{N_s} \left(\frac{\partial A(m, \partial_t, \partial_x)}{\partial m} u^s, \lambda^s \right), \tag{25}$$

where $(.,.)$ is the scalar product in the wavefield space and $\lambda^s(x,t)$ is the solution of the adjoint state equation

$$A^T(m, \partial_t, \partial_x)\lambda(x,t) = g^s(x,t), \quad x \in \mathbb{R}^d \times [0,T]. \tag{26}$$

In equation (26), the adjoint operator of the partial differential operator A is denoted by A^T. The source term $g^s(x,t)$ of the adjoint equation is given, after discretization, by

$$g^s = \sum_{r=1}^{N_r} R_r^T \left(\frac{\partial h_{L^p}(d_{cal}^{s,r}, d_{obs}^{s,r})}{d_{cal}} \right), \tag{27}$$

with

$$\frac{\partial h_{L^p}(d_{cal}, d_{obs})}{d_{cal,k}} = p|d_{cal,k} - d_{obs,k}|^{p-2}(d_{cal,k} - d_{obs,k}), \quad k = 1, \ldots, K. \tag{28}$$

Equations (25) to (28) can be interpreted as follows: the gradient of the misfit function $f_{L^p}[m]$ is given by the sum over the sources of a weighted zero-lag correlation between the incident wavefield $u^s(x,t)$ and the adjoint wavefield $\lambda^s(x,t)$. The weight depends on the derivatives of the partial differential operator $A(m, \partial_t, \partial_x)$ with respect to the parameter m. This term is known as the radiation pattern in the seismic imaging community. Most importantly, the source function for the adjoint computation is given by the derivative with respect to d_{cal} of the comparison function $h_{L^p}(d_{cal}, d_{obs})$ defined in (21).

In what follows we introduce a new FWI strategy where the comparison function is modified into a graph-space OT misfit function. Benefiting from the adjoint state formulation, computing the gradient of the corresponding misfit function requires only to modify the adjoint source as the derivative with respect to d_{cal} of the new comparison function. This result has been abundantly documented in the literature, thus we prefer not to describe its proof in this study and redirect the reader to reference papers such as the review from Plessix (2006) or the book from Chavent (2009) for more details.

3.2. Graph space optimal transport full waveform inversion

3.2.1. Mathematical development Introducing the vector $t \in \mathbb{R}^K$, $t = (t_1, \dots, t_K)$, the discrete graph of a seismic trace $d(t)$ is the ensemble of K points of $\mathbb{R}^2$ defined by $(t_1, d_1), \dots, (t_K, d_K)$. In short, we will denote this discrete graph by (t, d).

Using these notations, introducing the graph-space optimal transport misfit function amounts to modify the definition of the comparison function such that the new FWI problem is

$$\min_m f_{W^p}[m], \tag{29}$$

where

$$f_{W^p}[m] = \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} h_{W^p}(d_{cal}^{s,r}[m], d_{obs}^{s,r}), \tag{30}$$

with

$$h_{W^p}(d_{cal}, d_{obs}) = KW_p^p\Big((t, d_{cal}), (t, d_{obs}) \Big). \tag{31}$$

For simplicity of the following developments, we multiply the p-Wasserstein distance by the normalization factor K. Because (t, d_{cal}) and (t, d_{obs}) represent two point clouds in a 2D space, $h_{W^p}(d_{cal}, d_{obs})$ can be rewritten as the solution of the linear programming problem (16), where the points u_i and v_i now correspond to

$$u_i = (t_i, d_{cal,i}) \in \mathbb{R}^2, \quad v_i = (t_i, d_{obs,i}) \in \mathbb{R}^2. \tag{32}$$

Let σ^* be the permutation solution of the problem (16). We have

$$h_{W^p}(d_{cal}, d_{obs}) = \sum_{i=1}^{K} |t_i - t_{\sigma^*(i)}|^p + |d_{cal,i} - d_{obs,\sigma^*(i)}|^p \tag{33}$$

Solving the FWI problem (29) requires to access the gradient of the misfit function (30). Through the adjoint state technique, this only requires to express the derivatives of the comparison function $h_{W^p}(d_{cal}, d_{obs})$ with respect to d_{cal}. Here, we use the following

Theorem 1. *Let $X = (x_i)_{1 \le i \le K}$, $Y = (y_j)_{1 \le j \le K}$ be two ensembles of K points in $\mathbb{R}^d$, such that all the points y_j are distinct. Consider the function $F(X)$ which computes, for $p \ge 1$, the p-Wasserstein distance between the point clouds X and Y, defined by the LSAP*

$$F(X) = \min_{\sigma \in S(K)} \sum_{i=1}^{K} \|x_i - y_{\sigma(i)}\|^p, \tag{34}$$

where $\|.\|$ is the Euclidean norm in $\mathbb{R}^d$. Then, the solution $\sigma^*(X)$ to (34) is unique and locally constant a.e., to be defined in $\left(\mathbb{R}^d\right)^K$. Thus, $F(x)$ is differentiable a.e., and its gradient is given by

$$\nabla F(X) = p \begin{bmatrix} \|x_1 - y_{\sigma^*(1)}\|^{p-2} \left(x_1 - y_{\sigma^*(1)}\right) \\ \vdots \\ \|x_K - y_{\sigma^*(K)}\|^{p-2} \left(x_K - y_{\sigma^*(K)}\right) \end{bmatrix}. \tag{35}$$

In the specific case $p = 2$, we have

$$\nabla F(X) = 2\left(X - Y_{\sigma^*}\right) \tag{36}$$

where $Y_{\sigma^*} \in \left(\mathbb{R}^d\right)^K$ is defined by

$$\forall i, \ \ 1 \le i \le K, \ \ (Y_{\sigma^*})_i = y_{\sigma^*(i)}. \tag{37}$$

Proof. We start with the simpler case $p = 2$. In a second step, we generalize the proof to $p > 1$ and $p = 1$. We introduce the functions

$$F_\sigma(X) = \sum_{i=1}^{K} \|x_i - y_{\sigma(i)}\|^2 = \|X\|^2 - 2\,(X, Y_\sigma) + \|Y\|^2. \tag{38}$$

The LSAP (34) is thus equivalent to

$$\min_\sigma G_\sigma, \;\; G_\sigma(X) = -2\,(X, Y_\sigma)\,. \tag{39}$$

We have assumed that all the y_j are distinct, *i.e.*

$$\forall (j, j'), \;\; 1 \leq j \leq K, \;\; 1 \leq j' \leq K, \;\; j \neq j' \Longrightarrow y_j \neq y_{j'}. \tag{40}$$

Therefore, for $(\sigma, \sigma') \in S(K) \times S(K)$, we have

$$\begin{array}{rcl} Y_\sigma = Y_{\sigma'} & \Longleftrightarrow & \forall j, \;\; 1 \leq j \leq K, \;\; Y_{\sigma(j)} = Y_{\sigma'(j)} \\ & \Longleftrightarrow & \forall j, \;\; 1 \leq j \leq K, \;\; \sigma(j) = \sigma'(j) \\ & \Longleftrightarrow & \sigma = \sigma'. \end{array} \tag{41}$$

Therefore the Y_σ are all distinct. In addition, we have

$$G_\sigma(X) = G_{\sigma'}(X) \Longleftrightarrow (X, Y_\sigma - Y_{\sigma'}) = 0 \tag{42}$$

and the ensembles $H_{\sigma,\sigma'}$ such that

$$H_{\sigma,\sigma'} = \left\{ X \in \left(\mathbb{R}^d\right)^K, \;\; G_\sigma(X) = G_{\sigma'}(X) \right\}, \tag{43}$$

are hyperplanes of codimension 1, and therefore are null sets. The ensemble

$$H = \bigcup_{\sigma \neq \sigma'} H_{\sigma,\sigma'}, \tag{44}$$

is thus also a null set. For $X \notin H$, the functions $G_\sigma(X)$ are all distinct, and thus we have a unique minimizer

$$\sigma^*(X) = \arg\min_\sigma G_\sigma(X). \tag{45}$$

From the uniqueness of $\sigma^*(X)$ we have

$$\exists \delta > 0, \;\; \forall \sigma \in S(K), \;\; \sigma \neq \sigma^*(X), \;\; F_\sigma(X) - F_{\sigma^*}(X) > \delta. \tag{46}$$

Let $X' \in \left(\mathbb{R}^d\right)^K$ and consider

$$\sigma \in S(K), \;\; \sigma \neq \sigma^*. \tag{47}$$

By continuity of $F_\sigma(X)$ with respect to X, we have

$$\exists \eta_1 > 0, \;\; \|X - X'\| < \eta_1 \Longrightarrow F_\sigma(X') \geq F_\sigma(X) - \frac{\delta}{4} \tag{48}$$

and

$$\exists \eta_2 > 0, \;\; \|X - X'\| < \eta_2, \Longrightarrow F_{\sigma^*}(X) \geq F_{\sigma^*}(X') - \frac{\delta}{4} \tag{49}$$

Combining inequalities (46), (48) and (49) we obtain

$$\exists \eta = \min(\eta_1, \eta_2) > 0, \;\; \|X - X'\| < \eta \Longrightarrow \forall \sigma \in S(K), \;\; F_\sigma(X') > F_{\sigma^*}(X'). \tag{50}$$

This shows that $\sigma^*(X)$ is locally constant

$$\exists \eta > 0, \;\; \|X - X'\| < \eta, \;\; \sigma^*(X') = \sigma^*(X). \tag{51}$$

We have

$$\nabla F(X) = 2(X - Y_{\sigma^*}) + \frac{\partial F}{\partial \sigma} \frac{\partial \sigma^*}{\partial X}, \tag{52}$$

and from (51) we deduce that

$$\nabla F(X) = 2(X - Y_{\sigma^*}). \tag{53}$$

We now consider the case $p > 1$. We proceed in a similar way: we try first to characterize the ensemble of points X which satisfy, for $(\sigma, \sigma') \in S(K)^2, \;\; \sigma \neq \sigma'$

$$F_\sigma(X) = F_{\sigma'}(X). \tag{54}$$

For convenience, we introduce the notation

$$F_Y(X) = \|X - Y\|^p. \tag{55}$$

For two distinct point clouds $Y \in \left(\mathbb{R}^d\right)^K$ and $Z \in \left(\mathbb{R}^d\right)^K$, such that $Y \neq Z$, we have

$$\begin{aligned} \nabla F_Y(X) = \nabla F_Z(X) \quad \Longleftrightarrow \quad & \forall j, \;\; 1 \leq j \leq K, \\ & \|x_j - y_j\|^{p-2} (x_j - y_j) = \|x_j - z_j\|^{p-2} (x_j - z_j) . \end{aligned} \tag{56}$$

Let us consider one particular index k such that $y_k \neq z_k$. Dropping the index, we obtain

$$\|x - y\|^{p-2} (x - y) = \|x - z\|^{p-2} (x - z) \tag{57}$$

This implies that $x \neq y$, $x \neq z$, and that x, y, z are aligned along the same line. Up to a change of variables, we are in the situation where $y = 0$, $z = 1$, and $x = t$, such that it should satisfy

$$|t|^{p-2} t = |t - 1|^{p-2} (t - 1). \tag{58}$$

However this equation cannot be satisfied for $p > 1$. For $t \in [0, 1]$, the left hand side is negative and the right hand side positive. For $t > 1$, the left hand side is strictly greater than the right hand side. We reach symmetric conclusion for $t < 0$. Therefore, we conclude that

$$Y \neq Z \Longrightarrow \nabla F_Y(X) \neq \nabla F_Z(X). \tag{59}$$

From the assumption that the point y_j are all distinct, we have

$$\sigma \neq \sigma' \Longrightarrow Y_\sigma \neq Y_{\sigma'}. \tag{60}$$

Therefore, using (59), we have

$$\sigma \neq \sigma' \Longrightarrow \nabla F_\sigma(X) \neq \nabla F_{\sigma'}(X). \tag{61}$$

Consider now

$$G_{\sigma,\sigma'}(X) = F_\sigma(X) - F_{\sigma'}(X). \tag{62}$$

The ensembles $H_{\sigma,\sigma'}$, defined by

$$H_{\sigma,\sigma'} = \left\{ X \in \left(\mathbb{R}^d\right)^K , \;\; F_\sigma(X) = F_{\sigma'}(X) \right\}, \tag{63}$$

are characterized by

$$H_{\sigma,\sigma'} = \left\{ X \in \left(\mathbb{R}^d\right)^K , \;\; G_{\sigma,\sigma'}(X) = 0, \;\; \nabla G_{\sigma,\sigma'}(X) \neq 0 \right\}. \tag{64}$$

From the implicit functions theorem, (64) reveals that the codimension of $H_{\sigma,\sigma'}$ is 1. Therefore, (44) holds in the general case $p > 1$. The same conclusions as in the specific case $p = 2$ can thus be drawn: the functions $F_\sigma(X)$ are distinct a.e., and thus the minimizer $\sigma^*(X)$ is unique a.e. From the continuity of the functions $F_\sigma(X)$ with respect to X, we can show that the minimizer $\sigma^*(X)$ is locally constant, and thus obtain the general formula (35) for the gradient of $F(X)$.

Finally, the same conclusion can be obtained in the case $p = 1$. We first consider the ensembles on which the functions $F_Y(X)$ are not differentiable, defined by

$$\mathcal{E}_Y = \left\{ X \in \left(\mathbb{R}^d\right)^K, \;\; \exists j, \;\; x_j = y_j \right\}. \tag{65}$$

The ensemble $\mathcal{E}_Y$ contains a finite number of points and is therefore a null set.

For $Y \neq Z$, the equations (56) hold for $X \notin (\mathcal{E}_Y \cup \mathcal{E}_Z)$. We thus have,

$$\forall X \notin (\mathcal{E}_Y \cup \mathcal{E}_Z), \;\; \nabla F_Y(X) = \nabla F_Z(X) \Longleftrightarrow X \in \mathcal{D}_{Y,Z} \tag{66}$$

where $\mathcal{D}_{Y,Z}$ is defined by

$$\mathcal{D}_{Y,Z} = \{\exists j, \;\; x_j \in \text{Line}(y_j, z_j)\} = \bigcup_j \{X, \;\; x_j \in [y_j, z_j]\} \tag{67}$$

In (67), $\text{Line}(y_j, z_j)$ denotes the line connecting y_j and z_j in $\mathbb{R}^d$. Indeed, for $p = 1$, there exists a solution to the equations (56). However, the space $\mathcal{D}_{Y,Z}$ is of codimension 1. For two permutations $\sigma, \sigma' \in S(K)^2, \sigma \neq \sigma'$, we can thus define a space $H_{\sigma,\sigma'}$ of codimension 1, defined by

$$H_{\sigma,\sigma'} = \mathcal{E}_{Y_\sigma} \cup \mathcal{E}_{Y_{\sigma'}} \cup \mathcal{D}_{Y_\sigma, Y_{\sigma'}}, \tag{68}$$

such that

$$\forall X \in \left(\mathbb{R}^d\right)^K, \;\; X \notin H_{\sigma,\sigma'} \Longrightarrow F_\sigma(X) \neq F_{\sigma'}(X). \tag{69}$$

We can thus show, as for the previous case, that the solution to the problem (34) is unique and locally constant a.e.

□

The result from the theorem provides the gradient of the LSAP misfit function with respect to displacements of the point cloud X in all directions of $\mathbb{R}^d$. The FWI case we consider here is less general. We consider point clouds in $\mathbb{R}^2$, and we are interested in the sensitivity of the LSAP misfit function with respect to displacements of the point cloud X along one dimension only, the one corresponding to the amplitude (derivative with respect to d_{cal}). This means that we are interested in only a subset of the components of the gradient $\nabla F(X)$, those corresponding to variations along the amplitude axis. Following (35), we obtain

$$\frac{\partial h_{W^p}}{\partial d_{cal,k}} = p|d_{cal,k} - d_{obs,\sigma^*(k)}|^{p-2} \left(d_{cal,k} - d_{obs,\sigma^*(k)}\right). \tag{70}$$

3.2.2. Interpretation The cost $h_{L^p}(d_{cal}, d_{obs})$ given in equation (21) computes the distance between d_{cal} and d_{obs} by summing over local, sample by sample comparisons between $d_{cal,k}$ and $d_{obs,k}$. The corresponding adjoint source in equation (28) is the simple difference, sample by sample, between calculated data and observed data. This difference is weighted by its norm to the power $p - 2$, which indicates the well-known better resilience to outliers for p tending to 1.

Alternatively, the graph-space OT cost function $h_{W^p}(d_{cal}, d_{obs})$ in equation (33) computes the misfit between d_{cal} and d_{obs} through nonlocal comparisons: each sample i of the calculated data is assigned with a sample $\sigma(i)$ of the observed data through the solution of the graph-space OT problem. The comparison function h_{W^p} is the sum of the distance between samples from the calculated and observed data connected through this assignment. For each sample i, the time-delay $|t_i - t_{\sigma(i)}|^p$ associated with these two connected samples is added to the misfit measurement.

The corresponding residuals in equation (70) are also very similar to those from the L^p distance. Instead of being based on the ℓ^p residuals between $d_{cal,k}$ and $d_{obs,k}$, they correspond to the ℓ^p residuals between the sample k of the calculated data $d_{cal,k}$ and the sample $\sigma^*(k)$ of the observed data, also weighted by a power $p-2$ of this difference.

This comparison illustrates how the graph space OT misfit can be seen as a generalization of the L^p distance on $\mathbb{R}^K$: the local, sample by sample comparison, is replaced with a global comparison given by the optimal permutation computed through the graph-space OT problem.

Interestingly, we can actually show that the graph space OT misfit defines a distance in $\mathbb{R}^K$.

Theorem 2. *The misfit function* $\left(h_{W^p}(d_{cal}, d_{obs})\right)^{1/p}$ *defines a distance in* $\mathbb{R}^K$.

Proof. Let d_{cal}, d_{obs} be two vectors of $\mathbb{R}^K$.

Positivity. We have $h_{W^p}(d_{cal}, d_{obs}) \geq 0$.

Identity of indiscernibles. Assume that $h_{W^p}(d_{cal}, d_{obs}) = 0$. Then we have

$$\sum_{i=1}^{K} |t_i - t_{\sigma^*(i)}|^p + |d_{cal,i} - d_{obs,\sigma^*(i)}|^p = 0 \tag{71}$$

where σ^* is the permutation solution of the corresponding LSAP problem. Then

$$\forall i, \;\; 1 \leq i \leq K, \;\; t_i = t_{\sigma^*(i)}, \tag{72}$$

which means that $\sigma^* = I_d$ where I_d is the identity permutation of $\{1, \ldots, K\}$. Therefore

$$\forall i, \;\; 1 \leq i \leq K, \;\; d_{cal,i} = d_{obs,\sigma^*(i)} = d_{obs,i} \tag{73}$$

and $d_{cal} = d_{obs}$.

Symmetry. Let $f(\sigma)$ be such that

$$f(\sigma) = \sum_{i=1}^{K} |t_i - t_{\sigma(i)}|^p + |d_{cal,i} - d_{obs,\sigma(i)}|^p, \tag{74}$$

and $g(\sigma)$ such that

$$g(\sigma) = \sum_{i=1}^{K} |t_{\sigma(i)} - t_i|^p + |d_{cal,\sigma(i)} - d_{obs,i}|^p. \tag{75}$$

Introducing $j = \sigma(i)$ in the sum, g can be rewritten as

$$g(\sigma) = \sum_{j=1}^{K} |t_j - t_{\sigma^{-1}(j)}|^p + |d_{cal,j} - d_{obs,\sigma^{-1}(j)}|^p = f(\sigma^{-1}). \tag{76}$$

We introduce σ^* such that

$$h_{W^p}(d_{cal}, d_{obs}) = f(\sigma^*), \tag{77}$$

and $\widehat{\sigma}$ such that

$$h_{W^p}(d_{obs}, d_{cal}) = g(\widehat{\sigma}). \tag{78}$$

By definition we have,

$$\forall \sigma \in S(K), \;\; g(\widehat{\sigma}) \leq g(\sigma). \tag{79}$$

Therefore, using (76), we have

$$\forall \sigma \in S(K), \;\; f((\widehat{\sigma})^{-1}) \leq f(\sigma^{-1}), \tag{80}$$

which means that

$$\forall \sigma \in S(K), \quad f((\widehat{\sigma})^{-1}) \leq f(\sigma). \tag{81}$$

This shows that

$$(\widehat{\sigma})^{-1} = \sigma^*. \tag{82}$$

Therefore, using again (76),

$$g(\widehat{\sigma}) = f(\sigma^*), \tag{83}$$

and

$$h_{W^p}(d_{obs}, d_{cal}) = h_{W^p}(d_{cal}, d_{obs}). \tag{84}$$

Triangle inequality. We simply rely on the Wasserstein distance definition. For any $(d_1, d_2, d_3) \in (\mathbb{R}^K)^3$ we have

$$\begin{aligned} (h_{W^p}(d_1, d_3))^{1/p} &= W_p\left((t, d_1), (t, d_3)\right) \\ &\leq W_p\left((t, d_1), (t, d_2)\right) + W_p\left((t, d_2), (t, d_3)\right) \\ &\leq \left(h_{W^p}(d_1, d_2)\right)^{1/p} + \left(h_{W^p}(d_2, d_3)\right)^{1/p}. \end{aligned} \tag{85}$$

□

3.3. Practical implementation: how to choose time and amplitude extremal values?

The aspect ratio between the time and amplitude axis used to represent the points cloud plays a major role in the solution of the graph space OT problem. Intuitively, if the amplitude limits are significantly larger than the time limits, the OT solution will favor displacement along the time axis. In the opposite regime, the OT solution will favor displacement along the amplitude axis and we might end up with a distance similar to a L^p distance. A balance thus needs to be found between the maximum amplitude and time variations. We first analyze the situation for a single trace. Then we discuss how to perform properly this balance in a shot gather configuration, where the relative amplitude of each trace, as an imprint of the Amplitude Versus Offset (AVO) response, is different.

3.3.1. Single trace normalization We introduce

$$A^+_{cal} = \max_{1\leq i\leq K} d_{cal,i}, \quad A^-_{cal} = \min_{1\leq i\leq K} d_{cal,i}, \tag{86}$$

and

$$A^+_{obs} = \max_{1\leq i\leq K} d_{obs,i}, \quad A^-_{obs} = \min_{1\leq i\leq K} d_{obs,i}. \tag{87}$$

We define A as

$$A = \max\left\{A^+_{cal} - A^-_{cal}, A^+_{cal} - A^-_{obs}, A^+_{obs} - A^-_{cal}, A^+_{obs} - A^-_{obs}\right\} \tag{88}$$

The quantity A is the maximum amplitude variation in both observed and calculated data. The maximum time variation is simply T.

We introduce a parameter η in the ground cost to rescale properly the amplitude values with respect to the time values, such that we consider the (scaled) misfit measurement

$$h_{W^p}(d_{cal}, d_{obs}) = \sum_{i=1}^{K} c^{\eta}_{i\sigma^*(i)}, \tag{89}$$

with

$$c^{\eta}_{ij}(d_{cal}, d_{obs}) = |t_i - t_j|^p + |\eta\left(d_{cal,i} - d_{obs,j}\right)|^p. \tag{90}$$

Accordingly, we have

$$\frac{\partial h_{W^p}}{\partial d_{cal,k}} = p\eta^p |d_{cal,k} - d_{obs,\sigma^*(k)}|^{p-2} \left(d_{cal,k} - d_{obs,\sigma^*(k)}\right). \tag{91}$$

We propose to calibrate the coefficient η based on the aspect ratio between the time and amplitude axis, such that

$$\eta = \frac{\Delta T}{A}. \tag{92}$$

The parameter ΔT is a control parameter yielding the possibility to adjust the sensitivity of the graph space OT distance with respect to time and amplitude shifts. For a given ΔT, it costs the same, for an OT point of view, to displace a Dirac mass along the whole amplitude axis than from 0 to ΔT. Therefore, ΔT can be interpreted as the maximum time shift one might expect to recover from OT when applying the graph space OT strategy. For practical applications, it is easy to estimate this maximum time shift in the initial model.

In a regime where we have $\Delta T << T$, the cost associated with displacement along the amplitude axis becomes negligible. Therefore, the solution of the graph space OT problem consists in mapping the observed data with the calculated data through displacement only along this axis. The associated permutation solution of the LSAP problem would thus be $\sigma^* = I_d$, and

$$h_{W^p}(d_{cal}, d_{obs}) = \sum_{i=1}^{K} \eta |d_{cal,i} - d_{obs,i}|^p, \tag{93}$$

In this case we see that the misfit function $\widetilde{h}(d_{cal}, d_{obs})$ becomes equivalent to the L^p distance. In the limit case where ΔT reaches 0, the misfit function $h_{W^p}(d_{cal}, d_{obs})$ becomes identically equal to 0.

On the opposite regime $\Delta T >> T$, the cost associated with displacement along the time axis becomes negligible. In the schematic (and quite specific) case where the calculated and observed data are only shifted in time (perfect amplitude prediction), the solution of the graph space OT problem is σ^* such that

$$\forall i, \;\; 1 \leq i \leq K, \;\; d_{cal,i} = d_{obs,\sigma^*(i)}, \tag{94}$$

and we have

$$h_{W^p}(d_{cal}, d_{obs}) = \sum_{i=1}^{K} |t_i - t_{\sigma^*(i)}|^p. \tag{95}$$

The misfit function becomes sensitive only to the time shifts between events. If the time shifts are the same for all the samples in time, the misfit function is convex with respect to this time shift.

However, in practice, the calculated and observed data both differ by time and amplitude shifts: the amplitude of the seismic events is never perfectly predicted, both because of the assumptions in the numerical modeling and the presence of noise in the observations. As a consequence, for $\Delta T >> T$, the graph space OT misfit function would be dominated by the amplitude mismatch

$$h_{W^p}(d_{cal}, d_{obs}) \simeq \sum_{i=1}^{K} \eta |d_{cal,i} - d_{obs,\sigma^*(i)}|^p. \tag{96}$$

An illustration of the influence of ΔT is provided in Figure 1, where a 3D view of the discrete graph of two Ricker functions is presented. The two functions are shifted in time, and the red Ricker is scaled by a factor 1.25 so that the amplitude of the two functions is different.

The assignment solution of the LSAP problem is represented by gray arrows connecting the points from the graph of the shifted Ricker (in blue) towards the points of the reference Ricker (in red), following this assignment. Depending on the value of ΔT, the assignment can completely change. For small values such that $\Delta T = 0.4$ s, the total time being equal to $T = 40$ s, the particles are moved only along the amplitude axis. For intermediate values such that $\Delta T = 4$ s, the particles are displaced both along the time and amplitude axis. For large values such that $\Delta T = 20$ s, the particles are moved only along the time axis.

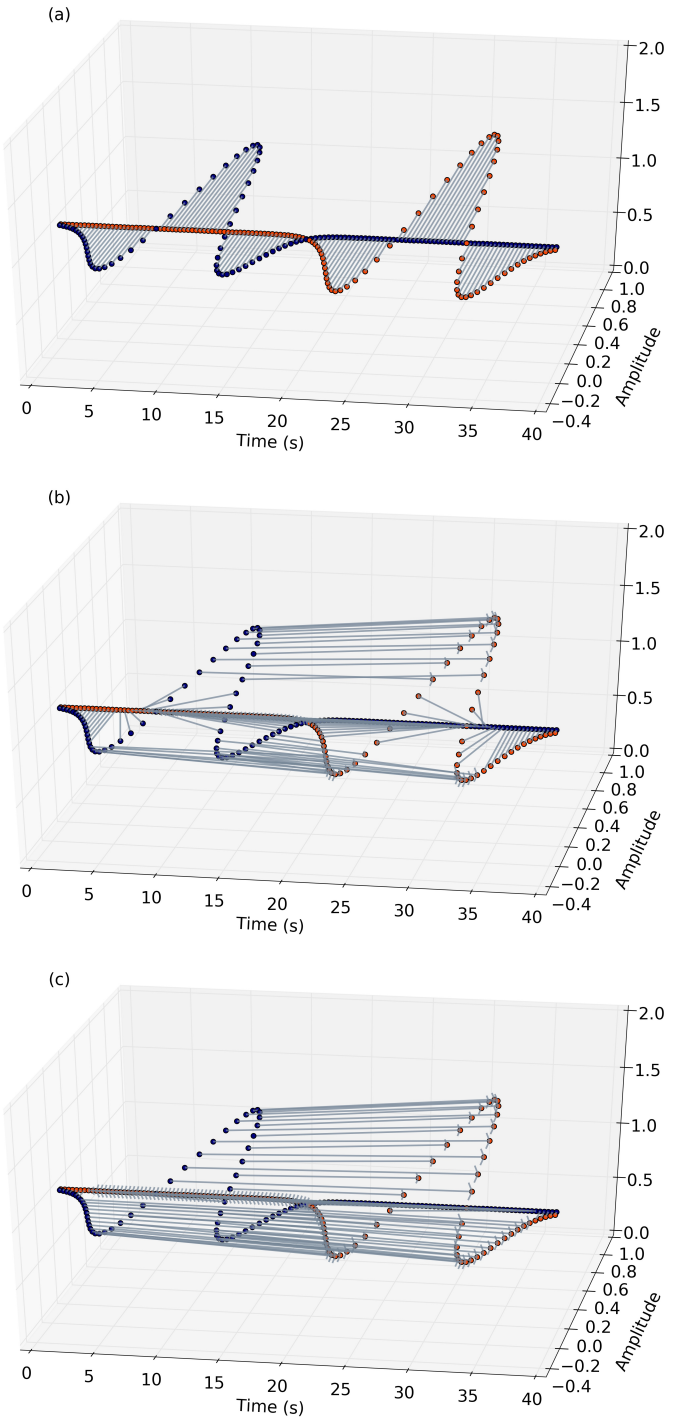

Figure 1: 3D representation of the discrete graph of a reference Ricker function (red points) and a shifted in time Ricker function (blue points) scaled in amplitude by a factor 0.8. The gray arrow represent the assignment solution of the LSAP problem, which depends on the value of the parameter ΔT. Top $\Delta T = 0.4$ s, middle $\Delta T = 4$ s, bottom $\Delta T = 20$ s.

3.3.2. Shot gather configuration Here we consider an observed and a calculated shot gather, defined as ensembles of N_r observed and calculated traces

$$d_{cal} = \left\{ d^1_{cal}(t), \ldots, d^{N_r}_{cal}(t) \right\}, \quad d_{obs} = \left\{ d^1_{obs}(t), \ldots, d^{N_r}_{obs}(t) \right\}. \tag{97}$$

For each trace, we define the quantity A^r (through 88). We define a single control parameter ΔT to be applied to each trace. However, as A_r depends on each trace, the normalization process amounts to a trace by trace scaling: the weight of each trace in the misfit function $f_{W^p}[m]$ becomes the same. This can be problematic, as the mean amplitude of each trace in a seismogram varies significantly. This variation is related to the Amplitude Versus Offset (AVO) effect, which is very important to take into account in the inversion for meaningful reconstruction of the reflectors (see Hampson, 1991, for instance). To restore the AVO signature in the misfit function, we propose to scale the contribution of each trace by the energy of the corresponding trace in the observed data. The misfit function that we finally consider is thus

$$f_{W^p}[m] = \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} w_{obs}^{s,r} h_{W^p}(d_{cal}^{s,r}[m], d_{obs}^{s,r}), \tag{98}$$

where the weights $w_{obs}^{s,r} \in \mathbb{R}_+$ are given by

$$w_{obs}^{s,r} = \frac{1}{T} \int_0^T |d_{obs}^{s,r}(t)|^2 dt. \tag{99}$$

3.4. Misfit function profiles

We start here the numerical investigations. To this purpose, we focus here, and in all the remainder of this study, on the W^2 case. This is a pragmatical choice, as we have observed difficulties to converge in the case $p = 1$. We have not investigated intermediate choices $1 < p < 2$. This might be the purpose of further studies. The details about the numerical strategy we employ to solve the LSAP problems are given in Section 4. For now, we focus on illustrating some properties of the misfit function f_{W^p} and its gradient.

3.4.1. 1D Ricker case It is important to investigate how the graph space OT misfit function depends on the parameter ΔT. A first illustration based on shifted in time Ricker functions is proposed here. A reference Ricker function with an additive Gaussian noise is first computed (Fig. 2). The graph space OT misfit depending on the time shift between this reference Ricker function and a synthetic Ricker function is computed for different values of ΔT (Fig.3). The presence of noise makes impossible to reduce the mismatch to 0 between the Ricker functions even when for 0 time shift.

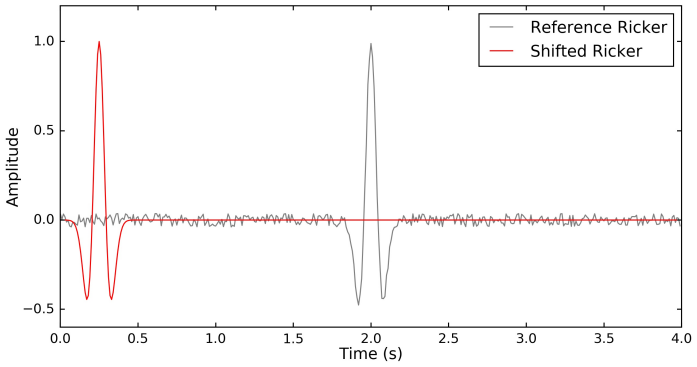

Figure 2: Reference Ricker function in solid gray line. Shifted in time Ricker function in solid black line.

As expected, when ΔT increases ($\Delta T = 40$ s), the impact of the amplitude mismatch dominates the misfit function and its profile depending on time shifts tends to a constant. On the opposite, when ΔT becomes small ($\Delta T = 0.04$ s) the solution of the OT graph space problem exhibits a profile similar to the least-squares misfit. For intermediate values of ΔT (around 0.4 s), the misfit function is sensitive to the amplitude mismatch and is convex with respect to the time shifts.

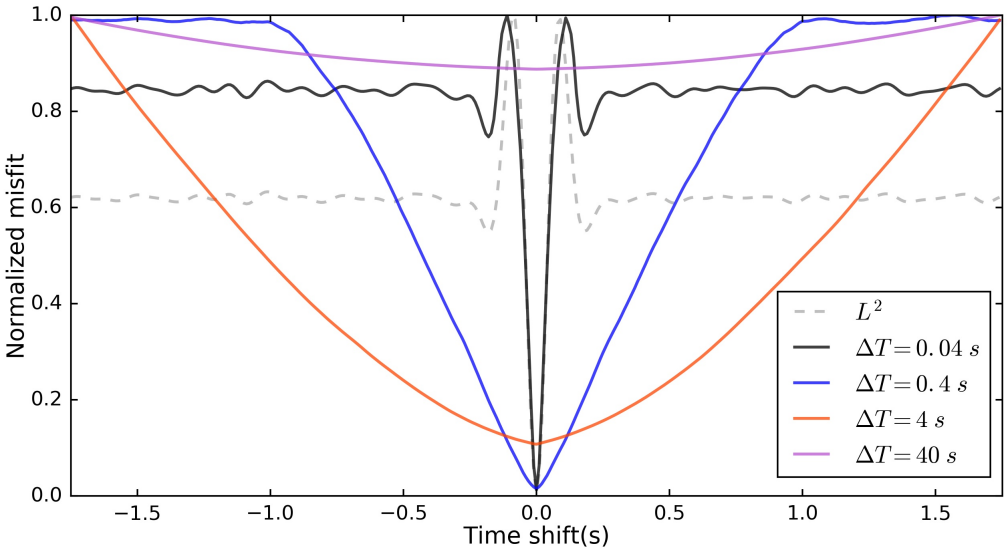

Figure 3: Comparison of the graph-space OT misfit function depending on the time shift, for different values of the parameter ΔT. The reference least-squares misfit is in dashed gray line. Solid black line: misfit function for $\Delta T = 0.04$ s. Solid blue line: misfit function for $\Delta T = 0.4$ s. Solid orange line: misfit function for $\Delta T = 4$ s. Solid purple line: misfit function for $\Delta T = 40$ s.

3.4.2. 2D misfit map We also reproduce an experiment initially proposed in Mulder and Plessix (2008), where a linearly increasing velocity model $v_P(z)$ is considered, such that

$$v_P(z) = v_{P,0} + \gamma z. \tag{100}$$

A 2D medium 3.5 km deep and 16.9 km long is considered. An initial data set is computed using references values $v_{P,0} = 2000$ m.s^{-1} and $\gamma = 0.7$ s^{-1}. A single source located just below the surface is used, at $x = 8.45$ km and $z = 0.05$ km. An array of 168 receivers is deployed at the same depth, from $x = 0.15$ to $x = 16.85$ km, with a regular spacing of 0.1 km. The time signature of the source is a Ricker function centered on 5 Hz, high-pass filtered to remove entirely the energy below 3 Hz. A Gaussian-Noise is added with a signal to noise ratio equal to 10. Misfit function maps are then computed for values of γ and $v_{P,0}$ in the range $\gamma \in [0.45, 0.95]$, $v_{P,0} \in [1750, 2250]$ with a sampling $\Delta\gamma = 0.0125$ s^{-1}, $\Delta v_{P,0} = 12.5$ m.s^{-1}.

We present first, in Figure 4, the observed data (black and white), together with the synthetic data in the model corresponding to $\gamma = 0.45$ and $v_{P,0} = 1750$ (blue and red). This allows estimating the maximum time shift expected with the slowest model for this experiment. For the largest offset, this time shift reaches 0.46 s.

The L^2 misfit function (Fig. 5a) is compared with the graph-space OT misfit function (98) for values of ΔT equal to $0.12, 0.23, 0.46$ and 2.3 s (Fig. 5b-e). As can be seen, the L^2 misfit function exhibits several local minima. When ΔT is small ($\Delta T = 0.12$ s, Fig. 5b), we still observe the presence of local minima. For $\Delta T = 0.23$ s, $\Delta T = 0.46$ s (Fig. 5c,d), the local minima eventually disappear, yielding smoother misfit functions. For $\Delta T = 2.3$ s, the misfit function becomes extremely flat, which is an indication that the amplitude mismatch starts dominating the time-mismatch, as is already observed in the 1D Ricker function test. The scaling with $\Delta = 0.46$ s seems to be an adequate choice: the misfit function has no local minima but sill with a narrow valley of attraction near the global minimum.

3.5. Adjoint source building: analogy with the cross-correlation misfit function

A better insight on the physical meaning of the adjoint source formula (70) can be obtained through the analysis of a simple transmission case study. We consider a single source/receiver acquisition in two 100 m deep boreholes, distant from 50 m. Both the source and the receiver are located

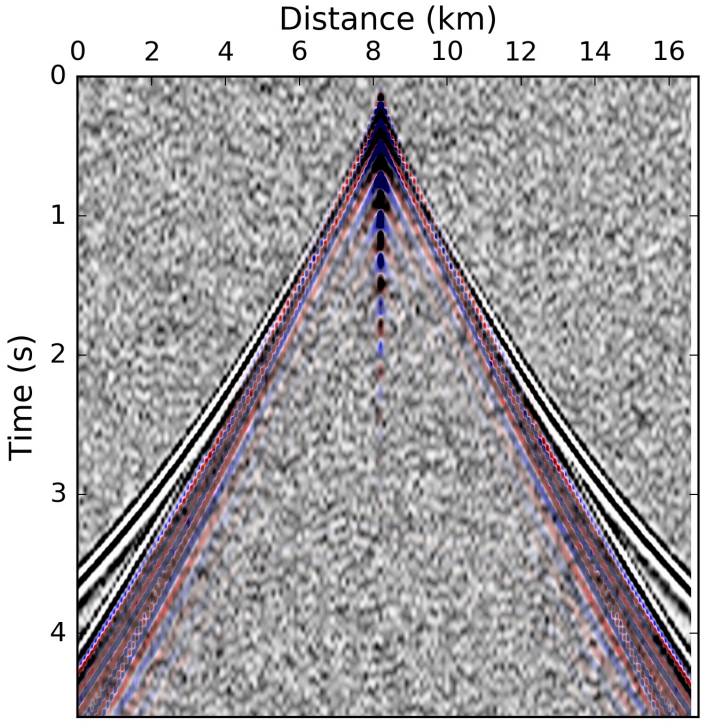

Figure 4: Observed data in the reference model with $\gamma = 0.7$ and $v_{P,0} = 2000$ m.s^{-1} (black and white). Synthetic data in the model with $\gamma = 0.45$ and $v_{P,0} = 1750$ m.s^{-1} (blue and red). We measure a maximum time shift at the largest offset equal to 0.46 s.

at 50 m depth. The source emits a Ricker pulse centered at 250 Hz. The receiver records the signal during 0.05 s. We consider a reference homogeneous velocity model at 2500 m.s^{-1}, a faster homogeneous model at 5000 m.s^{-1}, and a slower homogeneous model at 1500 m.s^{-1}. The three corresponding recorded signals are presented in Figure 6. The amplitude of the signal propagating in the slower medium arrives later and has a larger amplitude, while the signal propagating in the faster medium arrives earlier, with a smaller amplitude. The amplitude variations can be explained with the energy conservation principle.

We compute the adjoint source of the graph space OT misfit function for the faster and slower media. For the purpose of the analysis, we select a huge ΔT to enhance the sensitivity to time-shifts ($\Delta T = 2.5$ s). The resulting adjoint sources, normalized in amplitude, are presented in Figure 7. Interestingly, we see that the adjoint source corresponds to the observed data, shifted to the position of the calculated data, multiplied by the sign of the travel time difference between the observed and calculated data. In the case of a slower medium, the travel-time difference is positive, for a faster medium, this difference is negative, hence the change of polarity of the residuals in this case. This adjoint source shows strong similarities with the one associated with a well-known misfit function in the literature: the cross-correlation misfit function proposed by Luo and Schuster (1991). In their study, they introduce a tomographic misfit function based on the squared travel-times difference between observed and calculated data, where the travel time difference is estimated as the maximum of the cross-correlation of the observed and seismic trace. For simple single events traces, the method from Luo and Schuster (1991) provides a robust estimation of the traveltime difference and the misfit function is convex with respect to the time shifts.

It is interesting to note that the graph space OT approach exhibits a similar behavior in this case, while it is based on a completely different approach. The fact that the resulting adjoint source corresponds to the observed data shifted by the traveltime difference between observed and synthetic data can be explained by the formula (70). As ΔT is large, the assignment between the

calculated and observed discrete graph amounts to shift in time the observed data to the calculated. When building the adjoint source, for each time sample, we subtract to the calculated data the value of the observed data given by the assignment. In the case of a slower medium, the calculated data has a larger amplitude compared to the observed data, therefore the remaining function has still the sign of the original calculated data. In the case of a faster medium, the calculated data has a lower amplitude compared to the observed data, therefore the polarity of the source function changes.

The plot of the gradients in Figure 8 confirm the analysis. We can verify that the sign of the gradient in the first Fresnel zone depends on if the medium is slower or faster, which is the expected behavior for a tomographic misfit function.

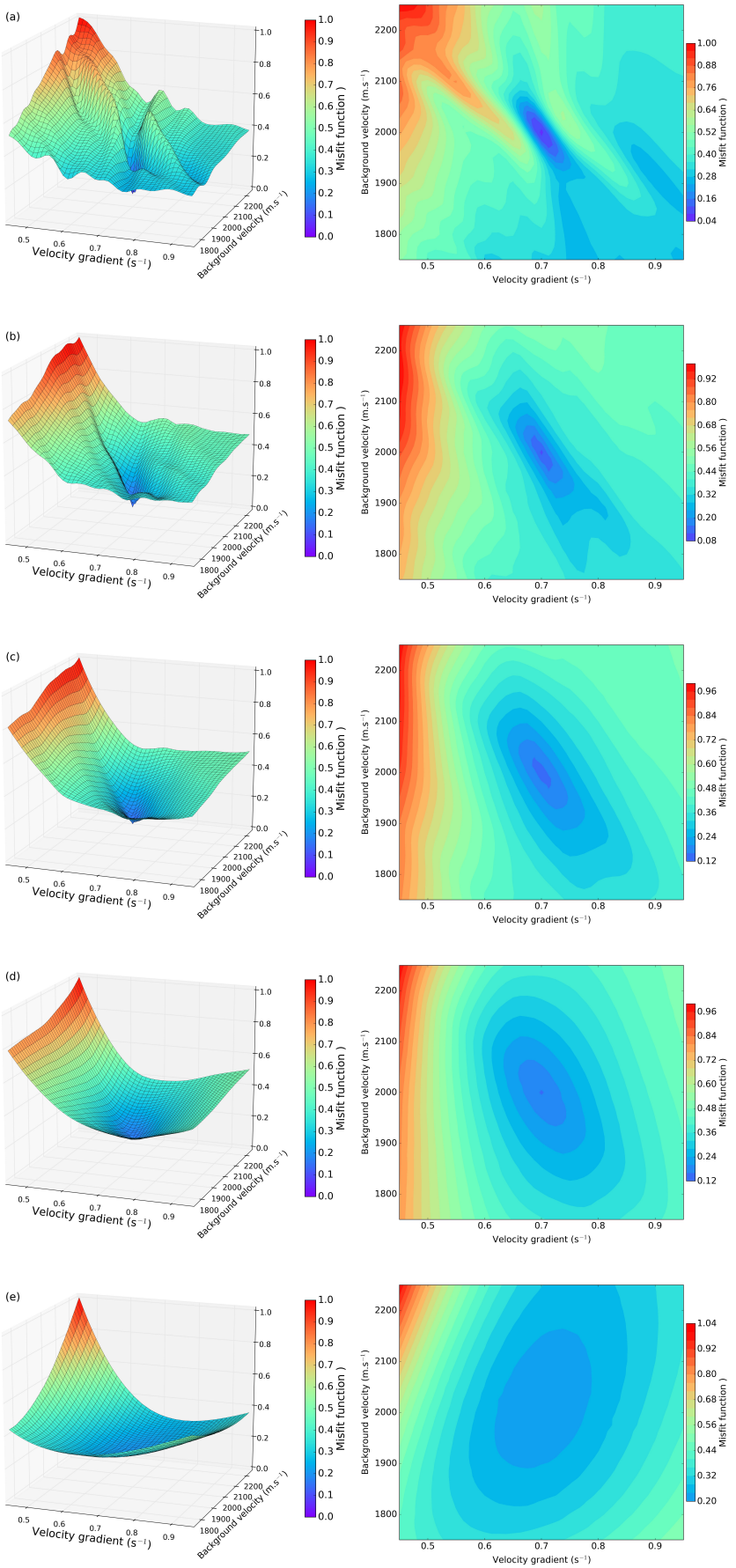

Figure 5: Misfit function maps for the linearly increasing velocity model. L^2 misfit function (a), graph-space OT misfit function for $\Delta T = 0.12$ s (b), $\Delta T = 0.23$ s (c), $\Delta T = 0.46$ s (d), $\Delta T = 2.3$ s (e).

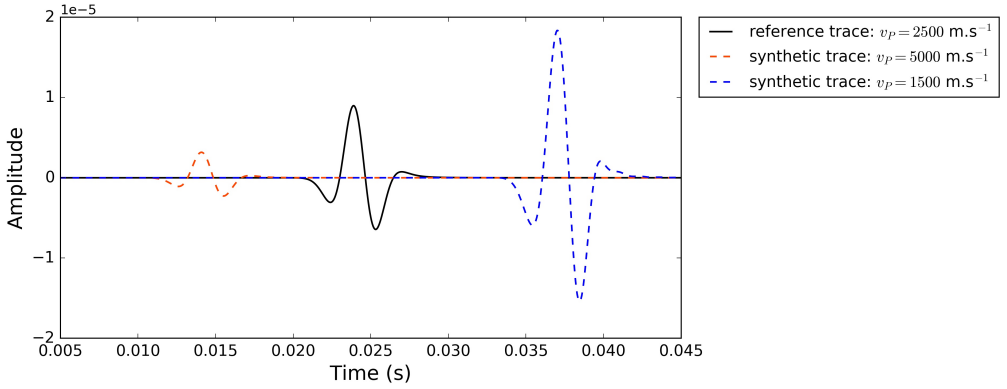

Figure 6: Signal recorded for the three considered homogeneous media: reference medium at 2500 m.s^{-1} (solid black line), slower medium at 1500 m.s^{-1} (dashed blue line), faster medium at 5000 m.s^{-1} (dashed orange line).

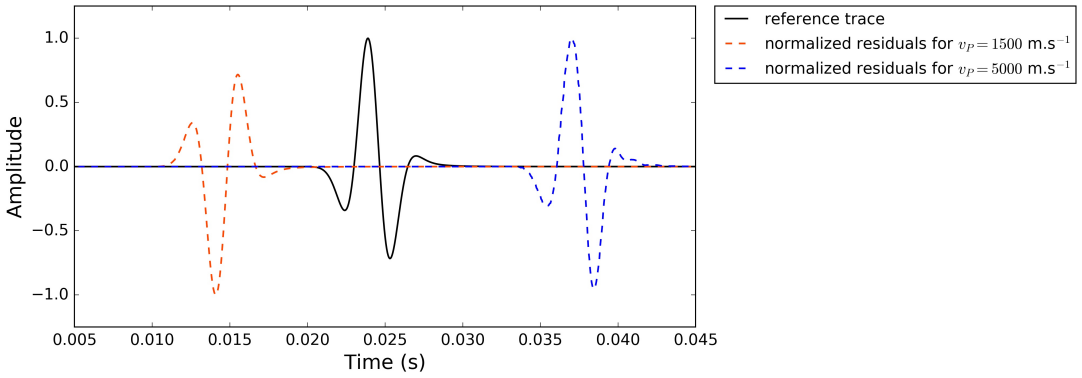

Figure 7: Reference signal for the medium at 2500 m.s^{-1} (solid black line). Corresponding graph space OT adjoint source for the slower medium at 1500 m.s^{-1} (dashed blue line), faster medium at 5000 m.s^{-1} (dashed orange line).

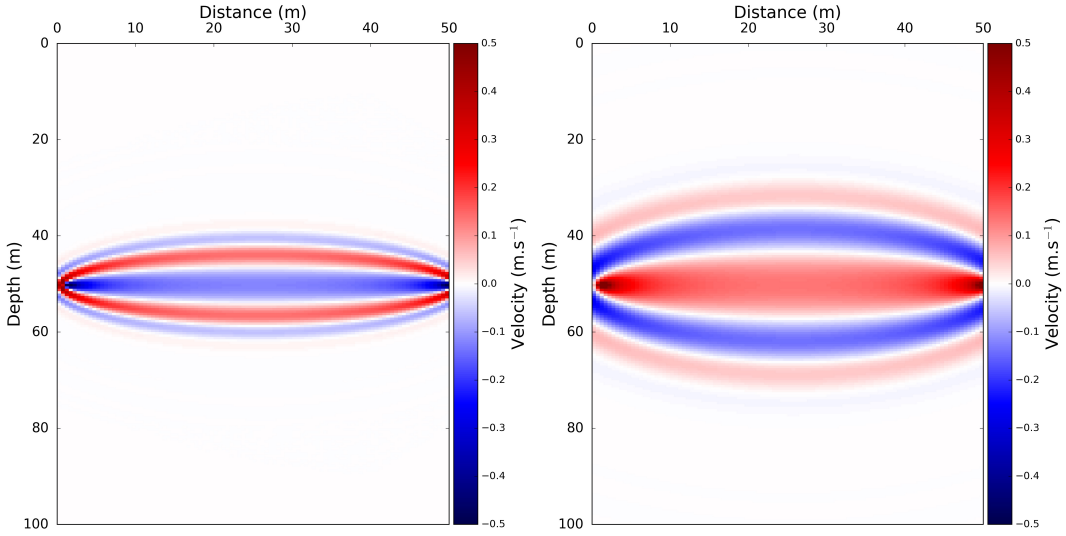

Figure 8: Gradient of the graph-space OT misfit function computed in the slower medium 1500 m.s^{-1} (left) and in the faster medium at 5000 m.s^{-1} (right).

4. Computing the Wasserstein distance between points cloud

The graph space OT misfit function exhibits interesting properties in terms of convexity for the FWI problem. In the perspectives of realistic size 2D and 3D FWI problems, we need to rely on an efficient numerical strategy to compute this misfit function. For realistic size acquisition systems, the number of sources and receivers can reach $O(10^3)$, leading to $O(10^6)$ seismic traces. For each of these seismic traces, a LSAP problem (16) needs to be solved. The size of this problem depends on the number of points required to discretize the seismic traces. Formally, this number should be associated with the Nyquist-Shannon sampling criterion: a minimum of two discretization points per period is necessary to correctly represent the signal. For realistic size FWI problems, the number of recorded periods is in the order of $O(10^2)$, leading to the same order for the size of the LSAP problem. As a consequence, implementing the graph-space OT approach presented in the last Section should lead to the solution of $O(10^6)$ LSAP problems of size $O(10^2)$ per iteration of the FWI local optimization solver, both to evaluate the misfit function and the adjoint source required to compute its gradient.

4.1. A short overview of existing methods

4.1.1. Generic OT solvers As the LSAP problem (16) originates from the solution of an OT problem, one might be interested in developments for the numerical solution of generic OT problems: the Benamou-Brenier approach based on a fluid dynamics interpretation of OT (Benamou and Brenier, 2000), methods based on the solution of the Monge-Ampère equations (Benamou et al., 2014), or solvers based on the dual formulation of the 1-Wasserstein distance (Métivier et al., 2016c). However, these solvers are not adapted to the specificity of the comparison of point clouds. They rely on a regular mesh discretization which is not satisfied by the point clouds: their repartition on the amplitude axis depends on the signal variation in time.

In Métivier et al. (2018), we circumvent this difficulty by representing the point cloud as a sum of Gaussian functions discretized over a 2D mesh. This leads to the solution of a 2D problem of size $O(KM)$, where M is the number of points associated with the discretization of the amplitude axis. A trade-off has to be found between a good representation of the signal, for which M is to be large, and a fast algorithm, which requires as small as possible M. In addition, as the Nyquist-Shannon sampling criterion assumes a floating point precision for the representation of the signal amplitude, using a too coarse amplitude discretization will require to increase the discretization in time as the 2 points per wavelength rule will not hold anymore. In practice, in the examples provided in (Métivier et al., 2018), it turns out that M is approximately in $O(K)$, with K corresponding to 5 points per wavelength. Using the W_1 solver developed in Métivier et al. (2016c), which has a linear complexity, this strategy ends up with a $O(K^2)$ complexity, and an observed computational times for one gradient 7 times larger than what is required for a conventional L^p misfit function, for a 2D FWI synthetic example. Besides the discretization, other tuning parameters need to be set, associated with the representation of the 2D point clouds as a sum of Gaussian functions. Although interesting for a proof of concept, this strategy seems difficulty suitable for realistic size problems.

Among generic OT solvers, two other options might be considered. The first is the entropic regularization approach, based on the introduction of an entropic regularization term in the standard OT problem (Benamou et al., 2015). The regularized problem can be solved using an iterative Bregman projection algorithm. The solution is alternatively projected onto two sets of constraints. Each projection requires a matrix-vector product involving a Toeplitz matrix when regular meshes are considered. As such, the method is quite powerful, each projection being done in $O(K \log K)$, and the convergence of the method is fast for sufficiently large regularization parameter. However, on irregular meshes, the matrix loses its Toeplitz structure. Each projection requires the multiplication of a vector by a dense matrix of size K, yielding a computational complexity in $O(K^2)$. Besides, preliminary tests we have performed show that for FWI applications, the regularization term needs to be relatively small, which requires the use of multi-level/multi-scale approaches. As entropic regularization is based on exponential functions, numerical stabilization steps need also to be enforced (Schmitzer, 2016).

Finally, semi-discrete strategies might be considered. This approach provides one of the most promising solvers for dealing with large scale instances of OT problems (Mérigot, 2011; Lévy, 2015). However, the method is not straightforward to adapt to the comparison of two point clouds as it relies on the assumption that one of the two probability measures which are compared can be described as a continuous function. In addition, the graph space OT approach considered in this study focuses on a large number of small scale dense problems, rather than on the solution of few large scale problems. For this reason, despite its efficiency for large scale instance OT problems, the multi-scale approach seems also not adapted to the needs of our application.

4.1.2. Linear programming solvers Instead of considering numerical OT solvers, the other options consists in considering linear programming solvers. Two classes of generic strategies exist: those based on the simplex algorithm (Dantzig, 1991) or those based on interior point methods (Karmarkar, 1984; Megiddo, 1989) (see for instance Gass (1984) for a review on linear programming algorithms). However, both methods do not fully exploit the specificity of LSAP problems of type (15).

Numerous economy problems can be modeled as LSAP problems. For this reason, numerous algorithms have been proposed for its solution during the second half of the twentieth century (see Bertsekas, 1998; Burkard et al., 2012, for a review). These algorithms can be divided in three main classes: those based on primal-dual methods (among them the Hungarian algorithm (Kuhn, 1955)); those based on a specification of the simplex algorithm to the solution of LSAP problems, either based on the primal (Akgül, 1993) or dual (Balinski, 1985) version of the simplex method; and purely dual algorithms, a category to which belongs the auction strategy introduced by Bertsekas and Castanon (1989). A benchmarking of all the existing LSAP solver for the cases we are interested in is beyond the scope of our investigation. From different studies (Bertsekas, 1998; Burkard et al., 2012), it appears that the auction algorithm, combined with an ε-scaling technique, achieves one of the best worst-case complexity among specific LSAP solvers. Benchmarking experiments on different sets of reference problems also show its good performance for the solution of small scale dense problems. Despite its theoretical cubic complexity, we therefore focus on the auction algorithm for our graph space OT strategy.

4.2. The auction algorithm

The auction algorithm is easier to interpret in the frame of maximum LSAP problem

$$\begin{aligned} \max_{\gamma_{ij}} \quad & \sum_{ij=1}^{K} \gamma_{ij} c_{ij}, \\ & \gamma_{ij} \geq 0, \qquad i,j = 1,\dots,K, \\ & \sum_{i=1}^{K} \gamma_{ij} = 1, \quad j = 1,\dots,K, \\ & \sum_{j=1}^{K} \gamma_{ij} = 1, \quad i = 1,\dots,K. \end{aligned} \tag{101}$$

We can always return to the minimization problem considering the identity

$$\max_x f(x) = -\min_x -f(x), \tag{102}$$

for a given function $f(x)$.

Considering a set of K persons $i = 1,\dots K$, a set of K objects $j = 1,\dots,K$, and a measure c_{ij} of the profit for the person i given by the object j, solving (101) amounts to assign to each person i a single object $j(i)$ such that the general profit $\sum_i c_{ij(i)}$ is maximized, and such that each object is assigned to a different person

$$i \neq k \Leftrightarrow j(i) \neq j(k). \tag{103}$$

The problem is easy to solve if, for each person i, the object $j_{max}(i)$ maximizing the profit defined by

$$j_{max}(i) = \arg\max_j c_{ij}, \tag{104}$$

is different from the objects maximizing the profit of the others

$$i \neq k \Leftrightarrow j_{max}(i) \neq j_{max}(k). \tag{105}$$

In this case the solution would be simply

$$\gamma_{ij} = \delta_{i,j_{max}(i)}, \tag{106}$$

where δ_{ij} is the Kronecker symbol such that

$$\delta_{ij} = \left| \begin{array}{ll} 1 & \text{if} \quad i = j, \\ 0 & \text{else.} \end{array} \right. \tag{107}$$

However, in the general case, the same object j might provide the maximum profit to more than a single person. A trade-off thus needs to be found between the persons interested in j to determine who should obtain it. This trade-off takes into account the profit provided by the other objects.

Consider the 2×2 example where c is such that

$$c = \begin{pmatrix} 10 & 1000 \\ 100 & 1000 \end{pmatrix}. \tag{108}$$

In this case the solution is

$$\gamma = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{109}$$

To maximize the general profit, the object 2 is assigned to the person 1: the second best object (object 1) provides him less profit than for to the person 2.

The trade-off to find between persons interested in the same object can be handled by assigning a price ψ_j to each object. This is the interpretation which can be given to the dual problem associated with (101), which can be written as

$$\min_{\psi_j} \sum_{i=1}^{K} r_i(v) + \sum_{j=1}^{K} \psi_j, \tag{110}$$

where $r_i(v)$ is the best possible profit for a person i given a set of prices ψ_j

$$r_i(v) = \max_{1 \leq jK} \left\{ c_{ij} - \psi_j \right\}, \quad 1 \leq i \leq K \tag{111}$$

This dual problem states that finding the assignment maximizing the general profit is equivalent to find a set of prices ψ_j as small as possible, which minimizes the sum of all individual profits r_i. Intuitively, in a competitive environment, if an individual profit is too large, it is detrimental to the profits of the others, which explains why the sum of all individual profits needs to be minimized.

The auction algorithm solves this dual problem. It is an iterative procedure based on an ε-relaxed version of the condition (111) written as

$$r_i = \max_{1 \leq j \leq K} \left\{ c_{ij} - \psi_j \right\} - \varepsilon, \quad 1 \leq i \leq K \tag{112}$$

The assignment γ_{ij} is initialized to 0, as well as the price vector ψ.

Each iteration starts with a bidding phase. A subset $I \subset 1, \ldots, K$ of non assigned persons is selected. For each person $i \in I$, the best φ_i^1 and second best profit φ_i^2 are computed

$$\begin{array}{l} j(i) = \arg\max_{1 \leq j \leq K} \left\{ c_{ij} - \psi_j \right\}, \\ \varphi_i^1 = c_{ij(i)} - \psi_j(i), \\ \varphi_i^2 = \max_{1 \leq j \leq K, j \neq j(i)} \left\{ c_{ij} - \psi_j \right\}. \end{array} \tag{113}$$

From these two values the bid $b_{ij(i)}$ proposed by i for object $j(i)$ is computed

$$b_{ij(i)} = \psi_j + \varphi_i^1 - \varphi_i^2 - \varepsilon. \tag{114}$$

The maximum increase of the price proposed by the person i to acquire $j(i)$ is the difference between the best profit and the second best profit for this person up to the relaxation parameter ε.

The second phase is the assignment phase. For each object j, we consider the set of persons $P(j)$ from which j has received a bid. If $P(j)$ is nonempty, the price ψ_j given to object j is increased by the best bid

$$\psi_j = \max_{i \in P(j)} b_{ij}. \tag{115}$$

Previously existing assignment implying j are canceled

$$\forall i = 1, \ldots, K, \;\; \gamma_{ij} = 0. \tag{116}$$

The object j is assigned to the current best bidder

$$i(j) = \arg\max_{i \in P(j)} b_{ij}, \;\; \gamma_{i(j)j} = 1 \tag{117}$$

The iterations continue until all the objects have been assigned. The termination of the algorithm is proved in Bertsekas (1998). Furthermore, it can be demonstrated that the computed assignment $\overline{\gamma}^{\varepsilon}$ resulting from this algorithm satisfies

$$\sum_{ij} \overline{\gamma}_{ij} c_{ij} - K\varepsilon \leq \sum_{ij} \overline{\gamma}_{ij}^{\varepsilon} c_{ij} \leq \sum_{ij} \overline{\gamma}_{ij} c_{ij} + K\varepsilon, \tag{118}$$

where $\overline{\gamma}$ denotes the solution of (101). A multiscale process is thus usually employed, where a cascade of relaxed problems are solved using decreasing values of ε. The full algorithm is summarized in Algorithm 1 in the Appendix A.

Note that there is a freedom in how to choose the subset I of unassigned persons at each bidding phase. The Gauss-Seidel version of the algorithm corresponds to the case where a single unassigned person is chosen, while the Jacobi version corresponds to the case where all unassigned persons are selected. Numerical tests we have performed indicate that the Gauss-Seidel version is more efficient, therefore we focus on this particular approach in this study.

4.3. Computational complexity

We have implemented a version of the Auction algorithm 1 in `FORTRAN`, which we have validated with a comparison with the GLPK simplex and interior point solvers (Makhorin, 2017). We compare the computational time obtained for the comparison of two point clouds associated with the discrete graph of two shifted Ricker functions. The time discretization varies from 200 to 6400 samples. The corresponding computational times are reported in the Table 1.

In Figure 9, a log-log plot of these computational times depending on the problem size K is provided, together with expected cubic and quadratic complexities from the same starting point ($K = 200$). As can be seen, the general trend is to follow a cubic complexity. However, for small size problems between 200 and 1000 the complexity is almost quadratic (estimated $O(N^{2.15})$), and the computational time is lower than 1 s. This is the range in which typical exploration scale case study will belong, with 100 up to 500 recorded periods per seismic trace. For instance, for a frequency content up to 20 Hz and 10 s recording, this would yield approximately 200 periods, which leads to $K = 400$ according to the Nyquist-Shannon sampling criterion.

K	200	300	400	600	800	1000	1200	1400	1600	2000	3200	6400
time (s)	8e-3	2.8e-2	4.8e-2	0.12	0.24	0.38	0.59	1.28	2.09	3.85	10.7	84.8

Table 1: Observed computational time in seconds for increasing problem sizes K. The computation has been performed using an INTEL core i7 processor at 2.9 GHz, with 32 GB of RAM.

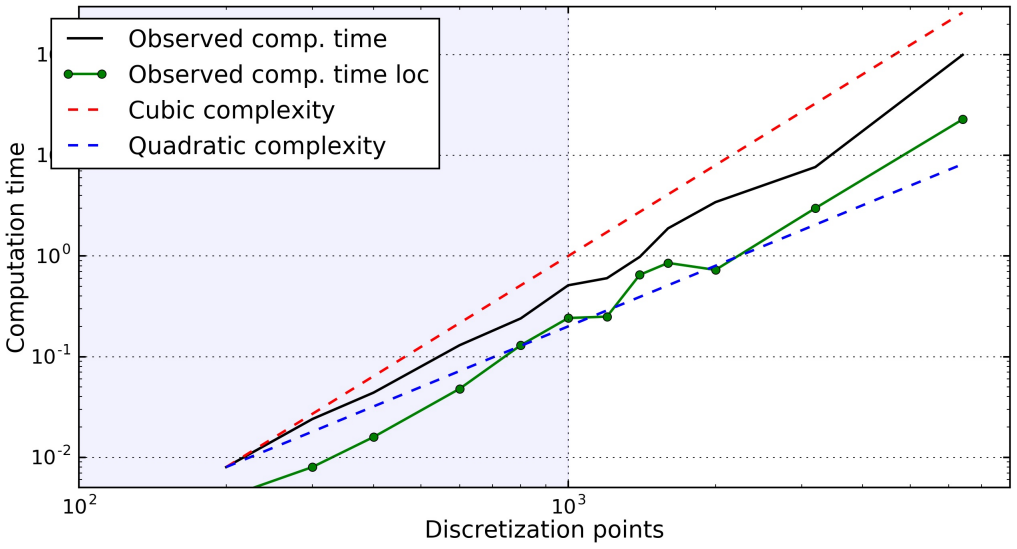

Figure 9: Observed computational complexity (black line) compared with theoretical cubic (red) and quadratic (blue) complexities in log-log plot. The shaded blue area focuses on the region of interest for standard size exploration case FWI case study. The computation has been performed using an INTEL core i7 processor at 2.9 GHz, with 32 GB of RAM.

5. 2D Valhall synthetic case study

To evaluate the interest and the feasibility of the graph space OT approach, we consider now the inversion of a 2D noisy synthetic dataset computed through a visco-acoustic modeling engine, with variable velocity, density and attenuation factor, and a signal to noise ratio equal to 10. From this data, only the velocity is reconstructed, the density and attenuation being considered as passive parameters. This challenging 2D case study is intended to be more realistic than experiments investigated in previous studies in inverse crime settings (for instance in Métivier et al., 2016c, 2018), where the amplitude of the seismic signal could be estimated with arbitrary accuracy. In this study, as the density and attenuation models used in the inversion are rough approximation of the correct ones, and also because of the noise, the amplitude cannot be predicted with perfect accuracy, as is always the case for the inversion of field data.

5.1. Presentation of the case study

We consider the 2D Valhall synthetic model. The Valhall oil field is a giant gas field located in the North Sea in a shallow water environment, which has been in production for several decades. Successful 3D FWI applications on the field data acquired in this area have led to the construction of accurate 3D velocity, density and attenuation models (quality factor) (Sirgue et al., 2010; Operto et al., 2015; Amestoy et al., 2016). These models, combined with local geological knowledge, have led us to the 2D synthetic models presented in Figure 10. The reservoir is located beneath the strong reflector between 2.5 and 3 km depth. On top of the reservoir, the presence of a layered gas cloud translates into low velocity horizontal layers, correlated with low density and low quality factor layers. In the first 500 m, the presence of unconsolidated shales also translates into a low velocity, low density, low quality factor zone, with thin horizontal layers.

From these 2D models, we first build a realistic synthetic dataset, based on the following visco-acoustic wave propagation model

$$\begin{cases} \partial_t v_x - \dfrac{1}{\rho}\partial_x p = 0, \\ \partial_t v_z - \dfrac{1}{\rho}\partial_z p = 0, \\ \partial_t p - \rho v_P^2(\partial_x u_x + \partial_z u_z) + \rho v_P^2 \displaystyle\sum_{l=1}^{L} Y_l \xi_l = \varphi, \\ \partial_t \xi_l + \omega_l \xi_l - \omega_l(\partial_x u_x + \partial_z u_z) = 0, \quad l = 1, \dots, L. \end{cases} \tag{119}$$

In (119), $v_x(x,z,t)$ and $v_z(x,z,t)$ are the horizontal and vertical velocity displacement, $p(x,z,t)$ is the pressure wavefield, $\varphi(x,z,t)$ is a pressure source. The P-wave velocity is denoted by $v_P(x,z)$ and the density by $\rho(x,z)$. The memory variables ξ_l implement the standard linear solid model to account for the attenuation. Following this model, the coefficients Y_l and the frequency ω_l are calibrated through a least-squares optimization to guarantee a frequency constant quality factor $Q(x,z)$ over a frequency band from 1 Hz to 30 Hz. Here we use three relaxation mechanisms, setting L to 3. Homogeneous initial conditions are considered. A free surface condition at the water/air interface is implemented on top. Sponge layers are used to mimic an infinite propagation medium at the bottom and on the lateral sides of the domain (Cerjan et al., 1985). A 4th order in space and 2nd order in time staggered grid finite-difference scheme is applied to discretize the system (119). More details on this modeling engine can be found in Yang et al. (2016a, 2018a).

A fixed spread acquisition system is used, with 128 sources at 50 m depth equally spaced each 130 m, and 169 receivers at 50 m depth equally spaced each 100 m. The sources are localized in space, such that for a given source location (x_s, z_s) we have

$$\varphi_s(x,z,t) = \delta(x - x_s)\delta(z - z_s)r(t). \tag{120}$$

The source time function $r(t)$ is designed as a Ricker function centered on 5 Hz, low-cut filtered so that the energy of the signal is equal to 0 below 2.5 Hz. Besides, a Gaussian white noise filtered in the frequency band $0 - 12.5$ Hz is added to each shot gather, with a signal to noise ratio equal

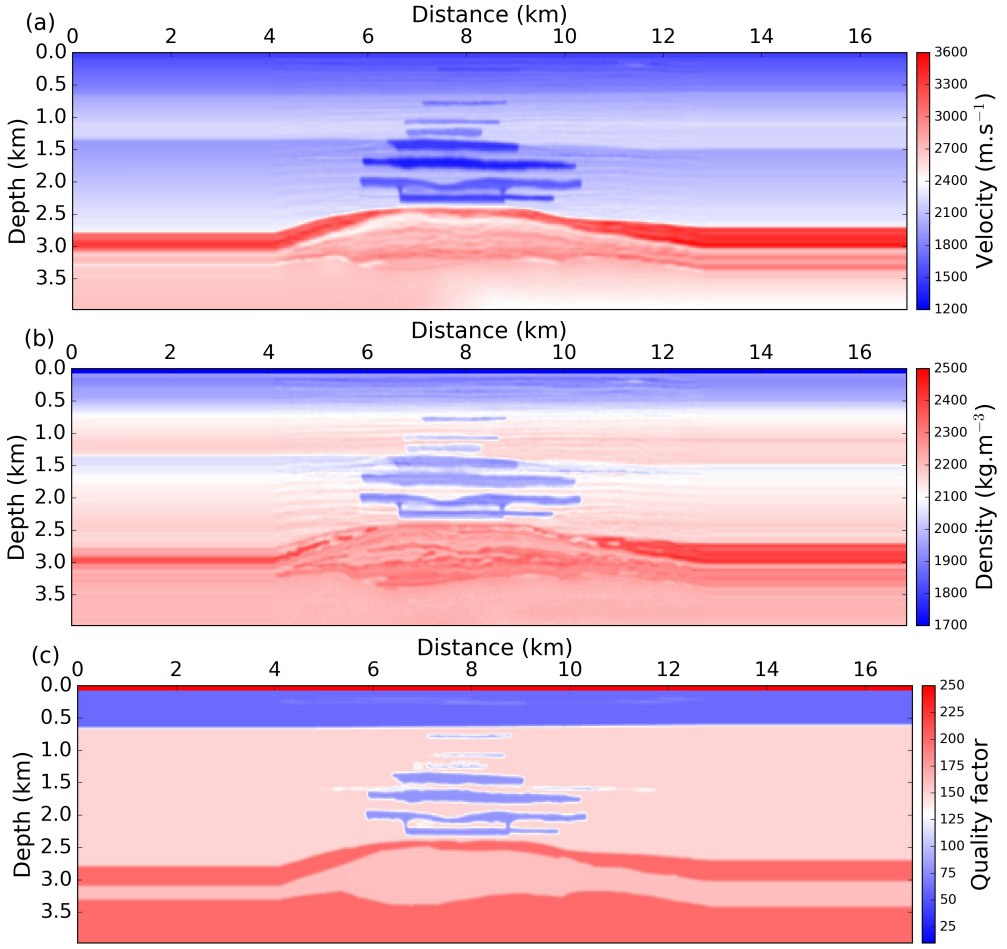

Figure 10: 2D Valhall model for velocity $V_p(x, z)$ (a), density $\rho(x, z)$(b), and quality factor $Q(x, z)$ (c).

to 10. The recording time is set to 8 s. The resulting 100th shot gather is displayed in Figure 11 as an example. An automatic gain control using a multiplication by the square root of the time is applied to enhance later arrivals, for visualization purpose only. As can be seen, the recorded signal is complex. Destructive interferences with the free surface remove the direct arrival around $x = 6$ km and $x = 12$ km. A strong reflection corresponding to the reflection on the interface at 2.75 km depth is recorded starting at 3.2 s for the receiver at $x = 13.5$ km. Diving arrival altered by their travel in the gas cloud regions are visible in the left bottom corner of the figure, from 4 to 8 s and from 0 to 8 km. The presence of noise tend to obscure all of these identifiable events.

The inversion is performed through a preconditioned l-BFGS algorithm, with a memory parameter l set to 20. The computation of the gradient is performed through a backpropagation of the incident wavefield from the boundaries to save memory usage, with a control on the wavefield reconstruction following the CARFS approach (Yang et al., 2016b). The whole gradient computation is parallelized over sources through MPI communication. The preconditioner approximates the diagonal of the Hessian operator through the pseudo-Hessian approach (Shin et al., 2001; Choi and Shin, 2008; Yang et al., 2018a). A regularization strategy based on a non-stationary smoothing of the gradient is implemented. The smoothing is Gaussian, with a correlation length equal to 1 time and 0.1 time the local wavelength in the horizontal and vertical directions respectively. The local wavelength is estimated through a reference frequency at 5 Hz and the local velocity value in the estimated velocity model.

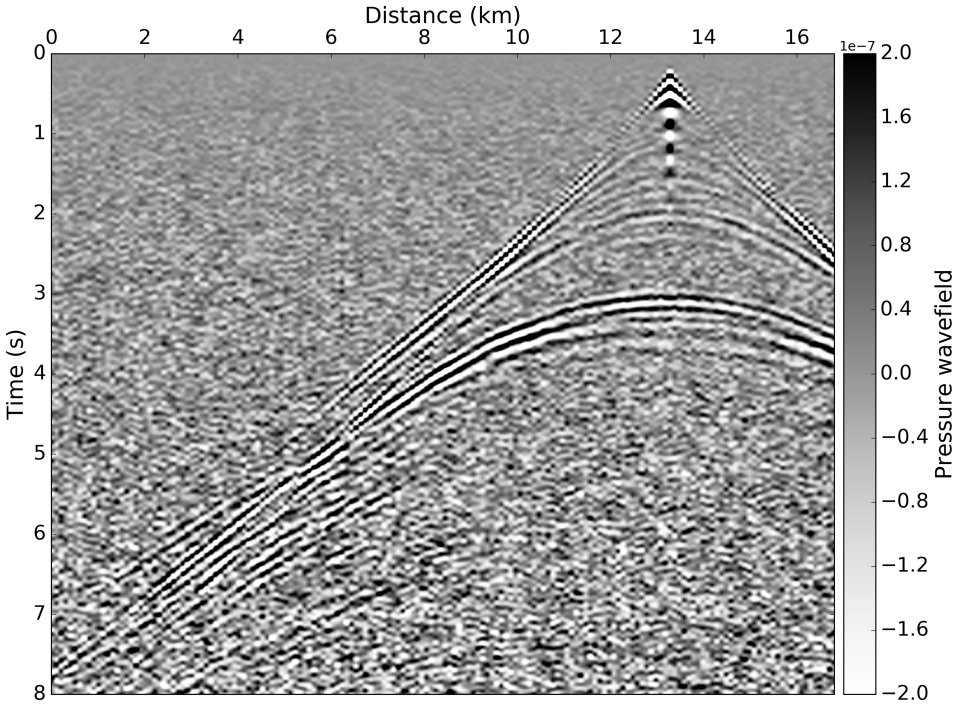

Figure 11: 100th shot-gather from the synthetic dataset. An automatic gain control (AGC) is applied to enhance the late arrivals for visualization purpose only. This AGC uses a multiplication by the square root of the time.

We consider three different initial models, derived from the exact P-wave velocity model, with different levels of smoothing. The first model is obtained through a Gaussian smoothing with a correlation length of 0.5 km, the second with a correlation length of 1 km, and the third with a correlation length of 1.25 km. For each initial model $v_{P,0}(x,z)$, a smooth density model $\rho_0(x,z)$ is derived using the Gardner law

$$\rho_0(x,z) = 1741 * (10^{-3} v_P(x,z))^{0.25}. \tag{121}$$

We also use a very simple input quality factor model $Q_0(x,z)$, which is equal to 1000 in the shallow water layer on top of the medium, and equal to 100 everywhere else.

To mimic the framework of a real data inversion, the source time signal $r(t)$ is estimated prior to inversion directly from the data. This is performed through a least-squares minimization in the frequency-domain, following the approach from Pratt (1999). The source time signal is estimated using the initial velocity, density and attenuation models and is not updated afterwards.

To assess the interest of the graph space strategy, the inversion results obtained through the minimization of the graph space OT misfit function are compared with the L^2 approach and the Kantorovich-Rubinstein misfit function introduced in Métivier et al. (2016c,b), which can be expressed as

$$f_{KR}(m) = \sum_{s=1}^{N_s} \max_{\varphi \in \mathrm{Lip}_1} \sum_{r=1}^{N_r} \int_0^T \varphi(x_r,t)\,(d_{cal}^{r,s}(t) - d_{obs}^{r,s}(t))\,dt \tag{122}$$

where Lip_1 is the space of 1-Lipschitz functions on (x_r,t). The latter strategy has shown interesting properties for mitigating the cycle skipping issue in FWI, and also yields the possibility to directly account for 2D shot-gathers through optimal transport, beyond the trace-by-trace comparison on which are based other current optimal transport implementations (Qiu et al., 2017; Yang and Engquist, 2018). The main drawback of the KR approach is the loss of convexity with respect to

large time shifts due to the direct computation of the W_1 distance for non-positive data through its dual formulation.

Finally, based on the comparison between observed and synthetic shot-gather using the different initial models, the maximum time shift is estimated to approximately 0.4 s. This is the value we use to set the control parameter ΔT.

5.2. Inversion results

For each initial velocity model, the models obtained using these three strategies (L^2, KR, graph space OT) and the corresponding data fit for the 100th shot gather are presented in Figures from 12 to 17. For each optimization method, the stopping criterion is based on a maximum number of iterations set to 500.

For the first initial model, close from the true velocity model, there is no cycle skipping, and the three methods produce a reasonably accurate velocity model estimation (Fig. 12). The data fit in the final model is similar for the three strategies: near offset reflections and far-offset diving waves are correctly interpreted (Fig. 13).

From the second starting model, FWI based on the L^2 misfit function is not able to converge to a correct estimation of the velocity model. There are strong low velocity artifacts on both sides of the gas layers, and a high velocity artifact appears at shallow depth in the middle of the first gas layer (Fig. 14a). In comparison, both the KR and graph-space OT approach converge towards a correct estimation of the velocity model (Fig. 14b,c). The data fit is instructive: cycle skipping affects later arrivals at large offset, between 4 and 7 s /0 and 6 km distance, in the initial model. The L^2 misfit function does not recover from this time shift in the initial model, while both the KR and graph space OT misfit function correctly interpret the data (Fig. 15).

Finally, from the third model, FWI based on both the L^2 and KR misfit function is not able to produce a correct velocity estimation. The corresponding results are affected by large low velocity artifacts on both sides of the gas layers. A large high velocity artifact is visible in the shallow part of the two models. Low velocity artifacts are also visible on both estimation just below the main reflector between 2.5 km and 3 km. The KR estimation seems worst than the L^2 one, as the lateral coherency of this deep reflector is lost (Fig. 16a,b). In comparison, the graph space OT estimation is closer to the true velocity model (Fig. 16c). The imprint of the low velocity artifacts on the sides of the gas layers is still visible but those artifacts are strongly attenuated. Only the deepest part below 3.5 km is not properly recovered, but this is consistent with the fact that very weak information from this zone reaches the data: reflections from there have been strongly attenuated and reach the noise level. The data misfit in the initial and final models shows that the data is strongly cycle skipped in the initial model: the arrivals at large offset, from 4 to 7 s and 0 to 6 km distance are predicted with more than one period shift in the initial model. The L^2 and KR misfit functions are not able to correct the model to put the data in phase, contrary to the graph space OT misfit function (Fig. 17).

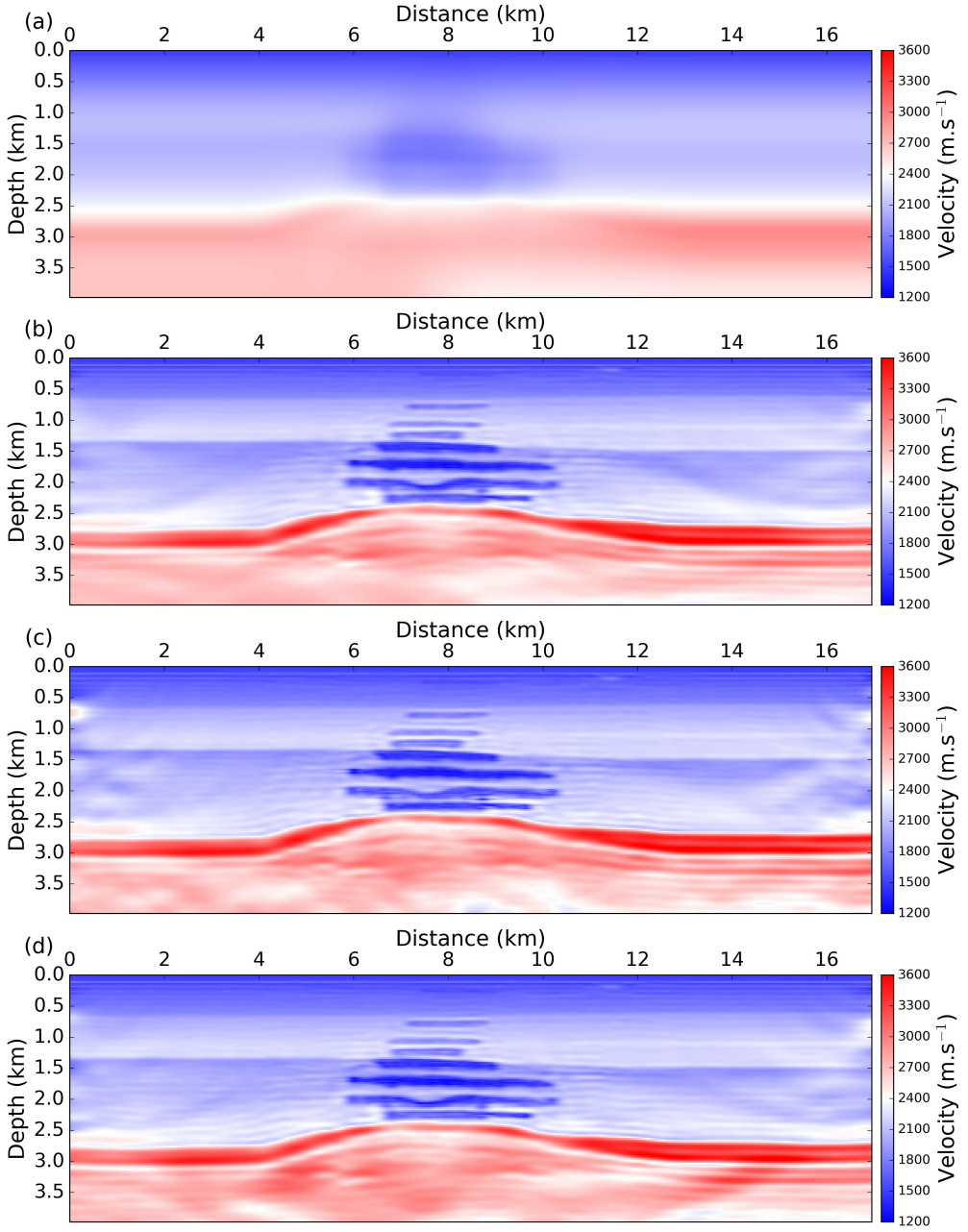

Figure 12: Initial model 1 (a), and corresponding reconstructed models using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function (d).

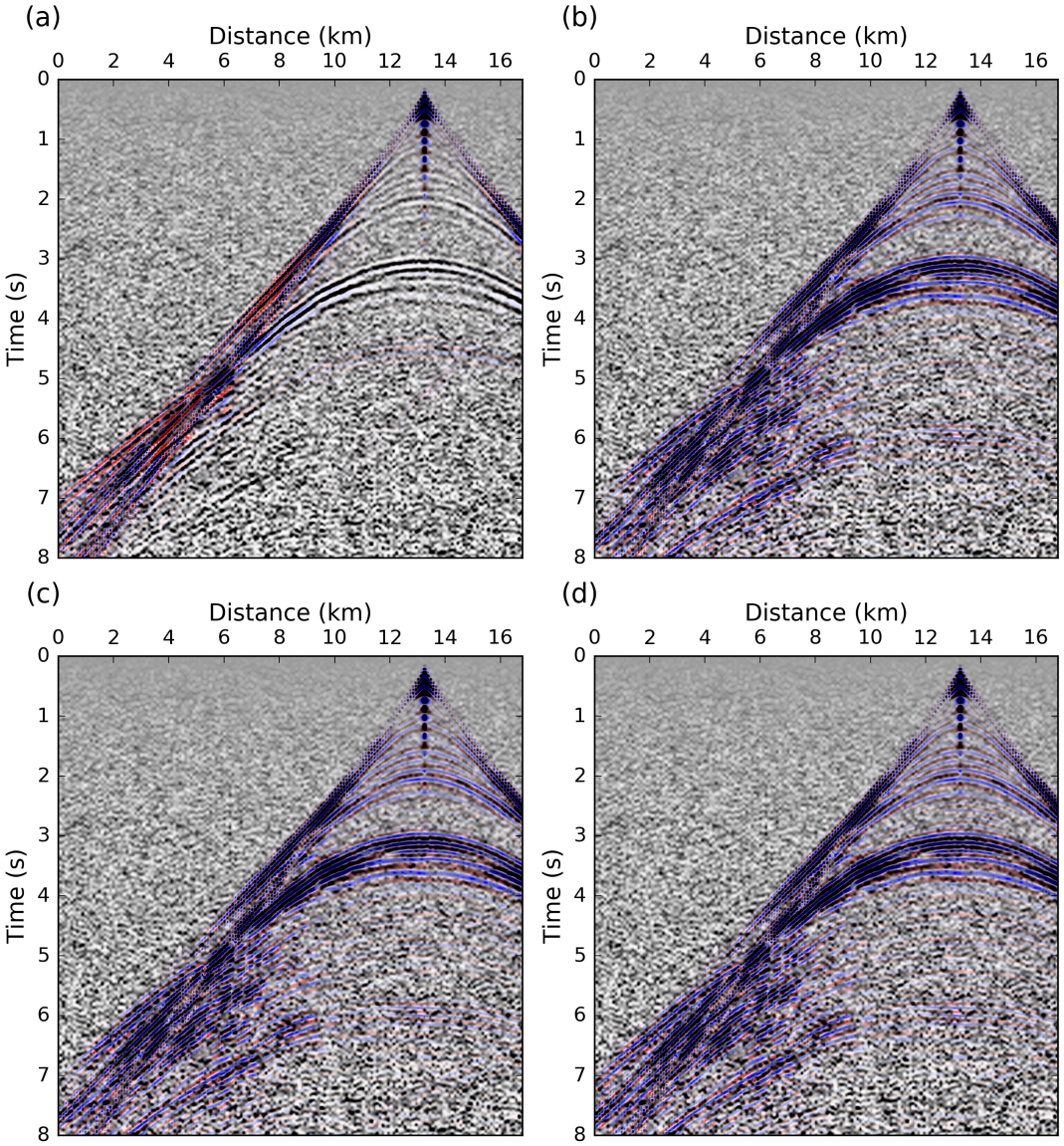

Figure 13: Data fit in the initial model 1 (a), and in the reconstructed model from model 1 using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function (d). The real data is represented in black and white, while the synthetic data computed in the corresponding models is represented in red and blue. When the synthetic data predicts correctly the true data, the blue overlays the white events, while the black overlays the red events.

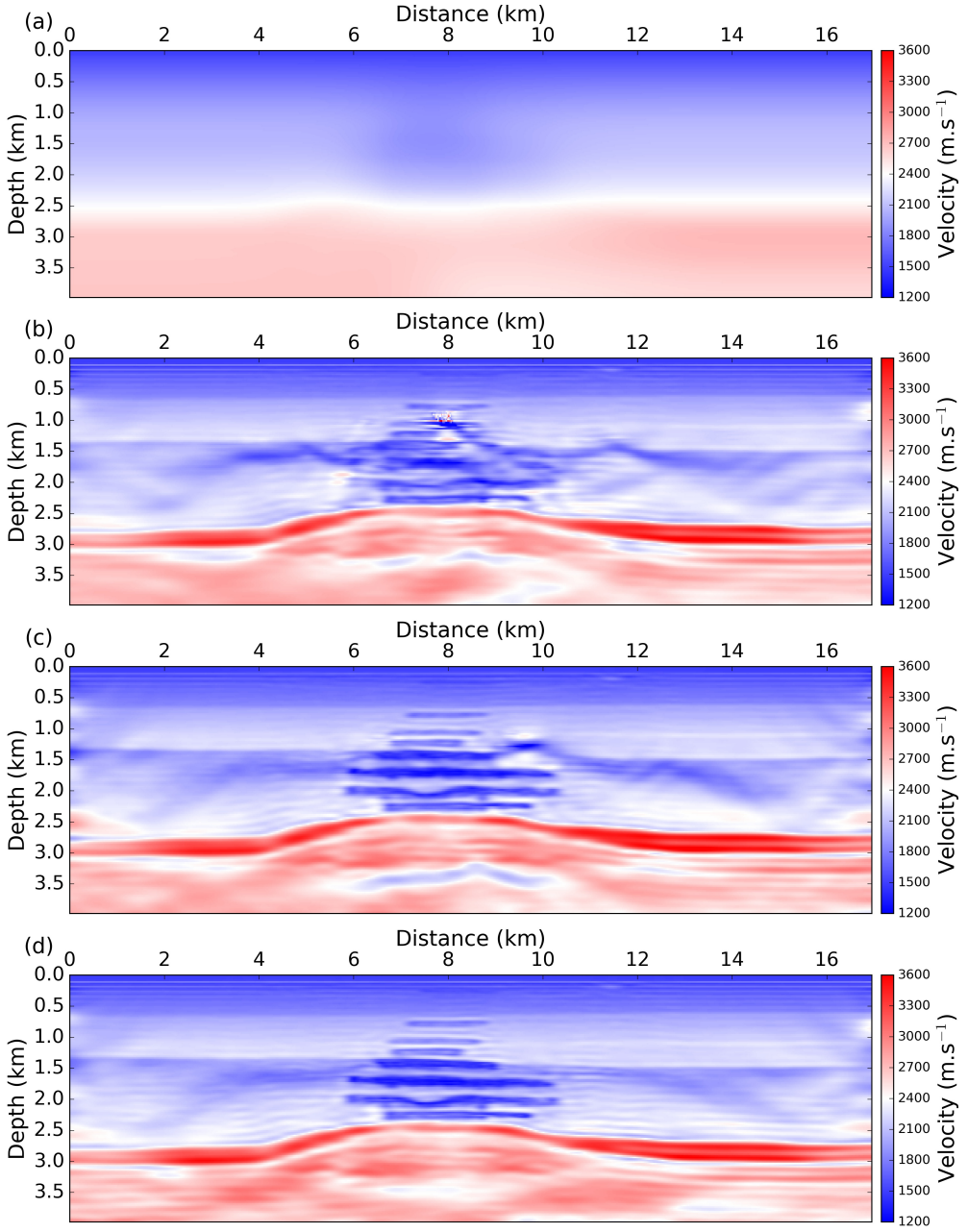

Figure 14: Initial model 2 (a), and corresponding reconstructed models using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function (d).

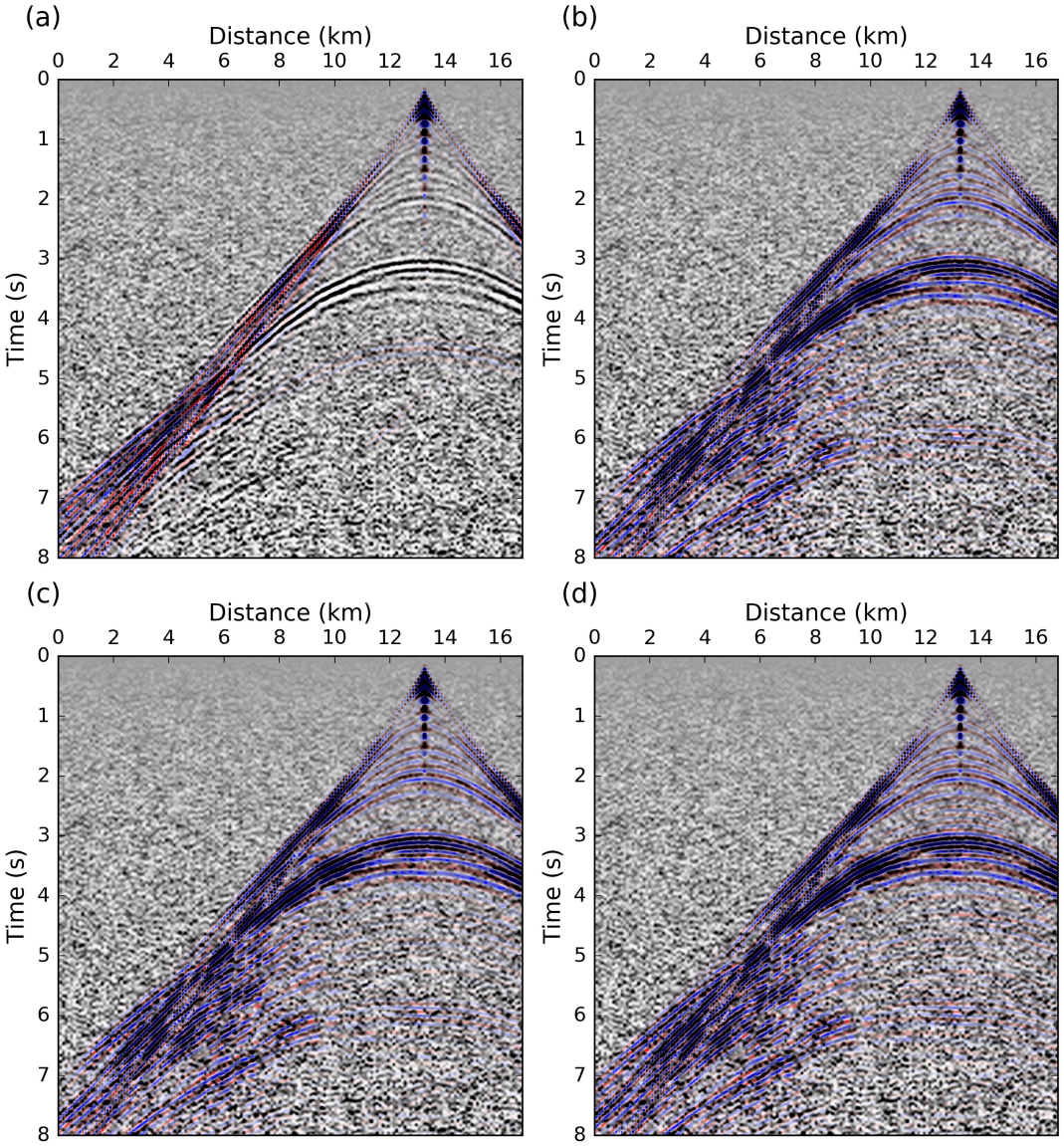

Figure 15: Data fit in the initial model 2 (a), and in the reconstructed model from model 2 using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function (d). The real data is represented in black and white, while the synthetic data computed in the corresponding models is represented in red and blue. When the synthetic data predicts correctly the true data, the blue overlays the white events, while the black overlays the red events.

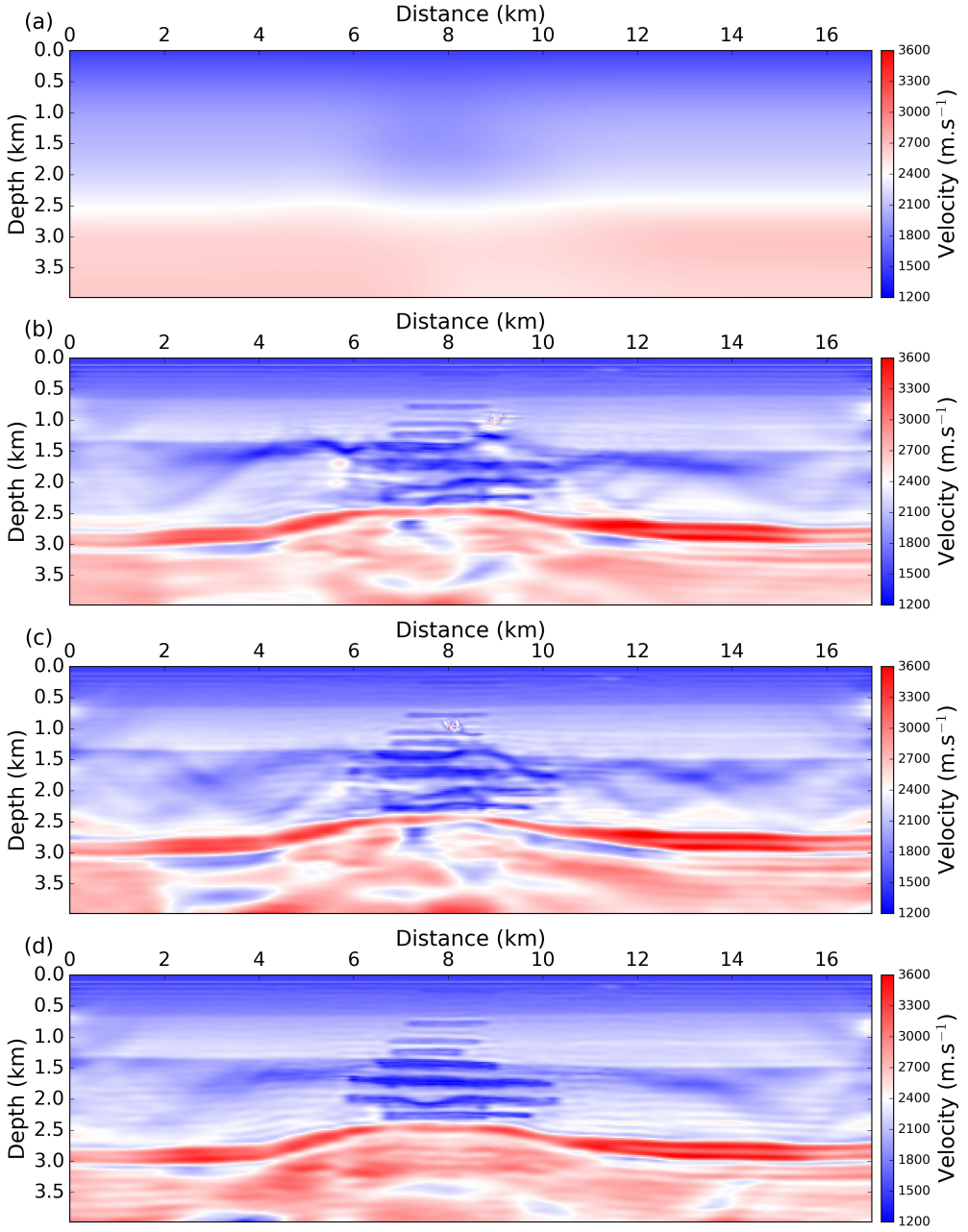

Figure 16: Initial model 3 (a), and corresponding reconstructed models using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function (d).

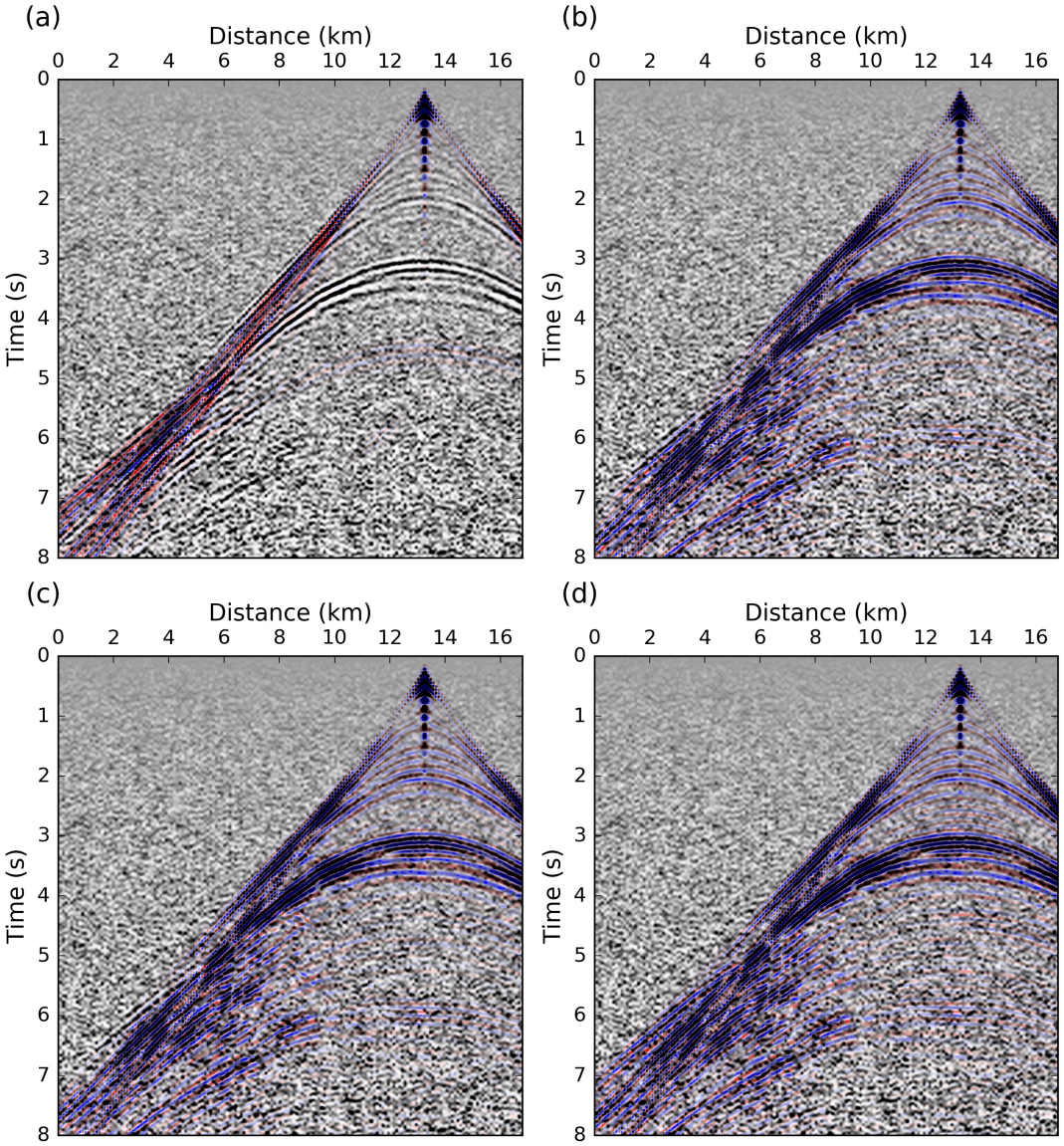

Figure 17: Data fit in the initial model 3 (a), and in the reconstructed model from model 3 using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function (d). The real data is represented in black and white, while the synthetic data computed in the corresponding models is represented in red and blue. When the synthetic data predicts correctly the true data, the blue overlays the white events, while the black overlays the red events.

5.3. Convergence analysis and computational cost

We focus on the third experiment to get a better insight on the convergence of the different methods in terms of model error and misfit reduction. The model error between an estimated model v_P and the exact velocity model v_P^* is defined as the average relative ℓ_1 error

$$e(v_P) = \frac{100}{M} \sum_{i=1}^{M} \frac{|v_{P,i} - v_{P,i}^*|}{v_{P,i}^*} \tag{123}$$

where M is the total number of discretization points for the velocity models. The corresponding convergence curves are presented in Figure 18. The misfit level is normalized therefore it starts at

1 for the three methods. It monotonically decreases along the iterations (Fig. 18a). For the three strategies, two regimes can be observed: a fast misfit decrease during the first 100 iterations, and then a slow misfit decrease. Interestingly, only the graph space OT approach produces a monotonic decrease of the model error (Fig. 18b). Both L^2 and KR starts increasing the model error. The lowest level of model error is reached by the graph space OT approach, confirming the qualitative observation made on Figure 16. The decrease of the model error with respect to the data error is also monotonic only for the graph space OT approach (Fig. 18c). The L^2 and KR approaches increase regularly the model error during the first phase of fast misfit decrease. The model error reduction is stronger during the second phase of slow misfit decrease, both for L^2 and graph space OT approaches.

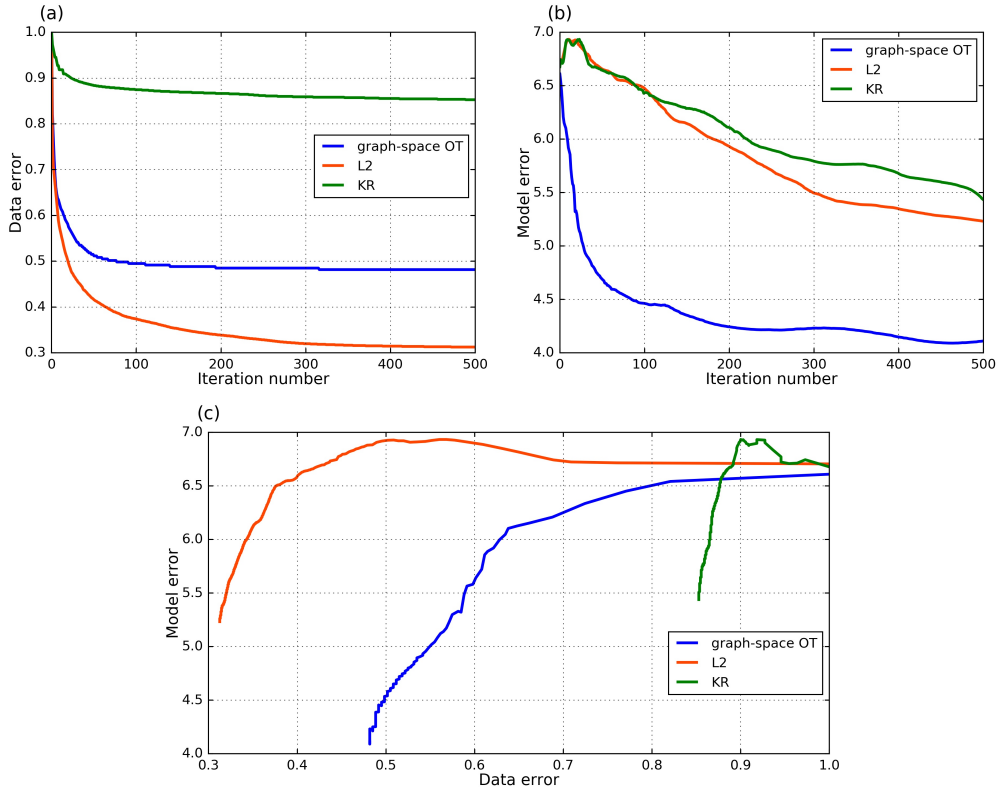

Figure 18: Convergence analysis for the inversion starting from the initial model 3. Misfit function decrease depending on the number of iterations (a). Model error depending on the number of iterations (b). Model error depending on the data error (c).

In terms of computational cost, the increase per iteration associated with the graph space OT approach is limited to approximately 6 % compared to the L^2 approach on this 2D Valhall case. The time discretization is driven by the CFL condition, which leads to 2666 discrete points per seismic traces. The data is re-sampled according to the Nyquist-Shannon criterion before computing the graph space OT misfit. In this case, we reduce by a factor 8 the number of time points, leading to the comparison of point clouds of size 333. The number of receivers is equal to 169. The total time for the computation of the graph space OT misfit function and its corresponding adjoint source is 4 s, which makes in average 0.02 s per instance of graph space OT problem, consistent with the complexity analysis in Figure 9. The total computational time for one gradient is approximately 60 s (10 s for the incident wavefield propagation, and 42 s for the adjoint and incident backpropagation from the boundaries), while approximately 56 s for the L^2 approach. The KR approach, in the settings chosen for this case study, requires approximately twice the same time for the misfit and adjoint source evaluation, up to 8 s. These results are summarized in Table 2.

To complement this analysis, we can observe that, for a given maximum frequency ω, the computational complexity for the solution of the wave equation is in $O(\omega^4)$, while the one for

	Inc. wfld	Adj. wfld	misfit & adjoint source	others	Total time	ratio
L^2 (s)	10.0	42.0	**0.4**	4.0	56.4	100 %
KR (s)	10.0	42.0	**8.0**	4.0	64.0	113 %
graph space OT (s)	10.0	42.0	**4.0**	4.0	60.0	106 %

Table 2: Computational cost analysis for one gradient on the 2D Valhall synthetic case study, depending on the chosen misfit function. The values are given in seconds (s). The computations have been performed on Intel Gold-6130 (Skylake) processors at 2.1 GHz. The cluster is composed of nodes of 16 processors with 192 GB RAM per node.

computing the graph-space OT misfit function would remain in $O(\omega^3)$, which is a favorable trend for larger scale/higher resolution applications. In addition, the graph-space OT misfit function computation is embarrassingly parallel as it is built as a summation over seismic traces of the graph-space OT cost. This makes the method efficient even for 3D large scale FWI applications. Preliminary results on 3D marine data containing around 25000 seismic traces per seismic source have been achieved, yielding computational overcost to the order of 20 percent in the first frequency band (up to 5 Hz).

5.4. Trace misfit evolution: graph-space OT approach in practice

To get a closer insight on how the graph-space OT approach works, we extract the seismic trace located at $x_r = 7$ km from the shot gather presented in Figure 11. We compare it with the corresponding synthetic trace in the initial model, in the models obtained at iteration 10 and 50, and in the final model in Figure 19. The reference trace is in black, while the synthetic traces are in blue. We also present the assignment solution of the graph space OT problem attached to these couple of traces, as orange arrows connecting the reference trace and the synthetic trace. As can be seen, in the initial model, cycle skipping is visible. The sensitivity of the assignment to the phase mismatch appears as arrows bending in the horizontal direction. Already in the model obtained at iteration 10, the largest amplitude events have been put in phase. Further iterations reconstruct the events before and after the largest amplitude train of events. Of course, this is only a partial view of the reconstruction, as this seismic trace is interpreted together with more than $20,000$ others. However this gives an illustration how the graph space OT approach efficiently reduces the phase mismatch.

5.5. Starting from a smoother medium: hierarchical approach

As is already noted in Métivier et al. (2018), the Figure 3 suggests a hierarchical/multi-stage approach where one would start the inversion with a large value for ΔT and progressively decrease it, restarting for each value an inversion starting from the model obtained with the last value of ΔT. We illustrate this possibility starting from an initial model 4, obtained through Gaussian smoothing with correlation lengths of 2 km in both horizontal and vertical direction.

The inversion is started with $\Delta T = 0.5$ s, to be consistent with the use of this smoother initial model. We then perform a series of inversion for ΔT equal to 0.4 s, 0.2 s, 0.8 s, 0.4 s and 0.2 s. For each stage, 500 FWI iterations are performed. The final model we eventually obtained is compared to the initial, L^2 and KR models in Figure 20, as well as the data fit in Figure 21. As can be seen, despite we start from a very simple model with a significant underestimation of the velocity increase at depth, it is possible, through this multi-scale approach, to reconstruct a quite meaningful velocity model. Conversely, L^2 and KR approaches converge towards non-informative local minima. The final data-fit show that most of the visible events in the reference shot gather are interpreted in the final velocity estimation.

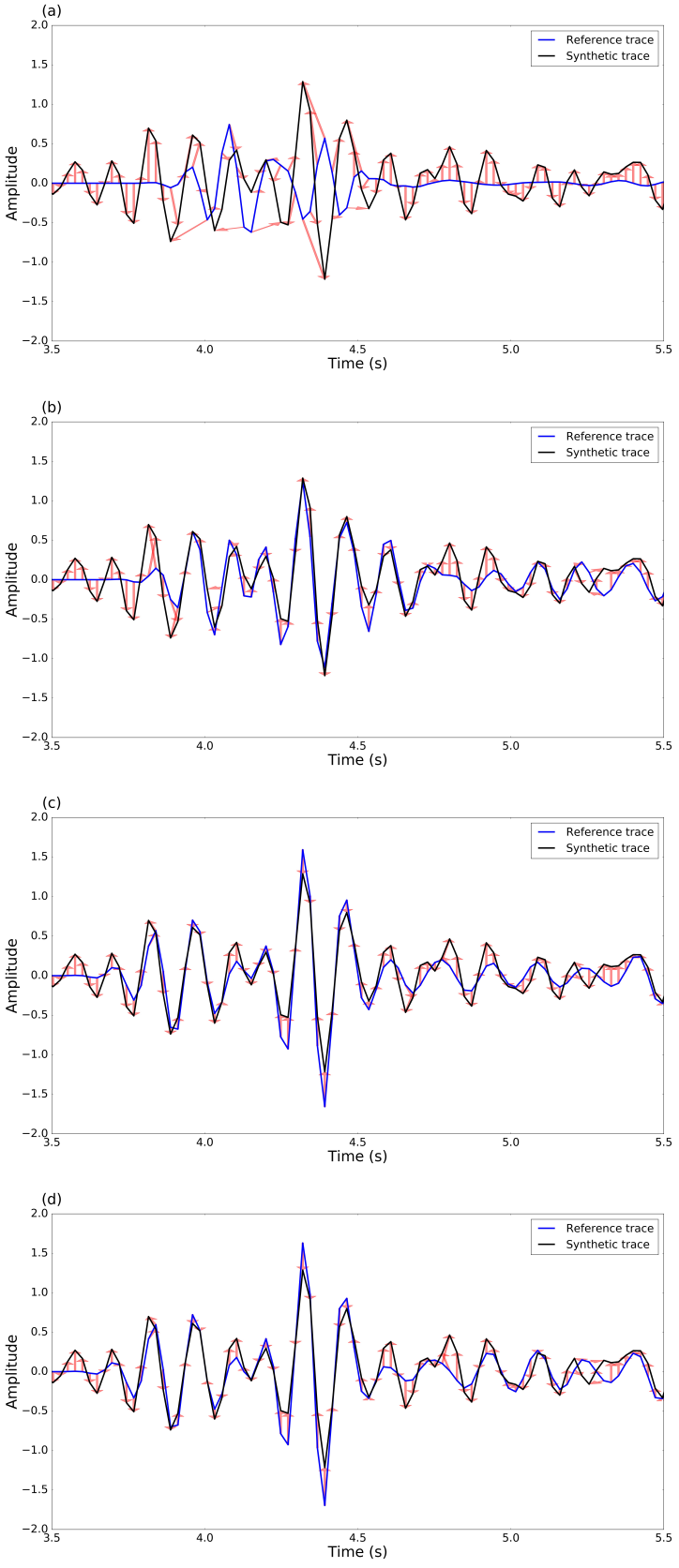

Figure 19: Reference seismic trace at $x_r = 7$ km (solid black line). Corresponding synthetic trace (solid blue line) in the initial model (a), 10th iteration model (b), 50th iteration model (c), final model (d). The assignment solution of the graph space OT problem between the synthetic and reference trace is presented in orange arrows.

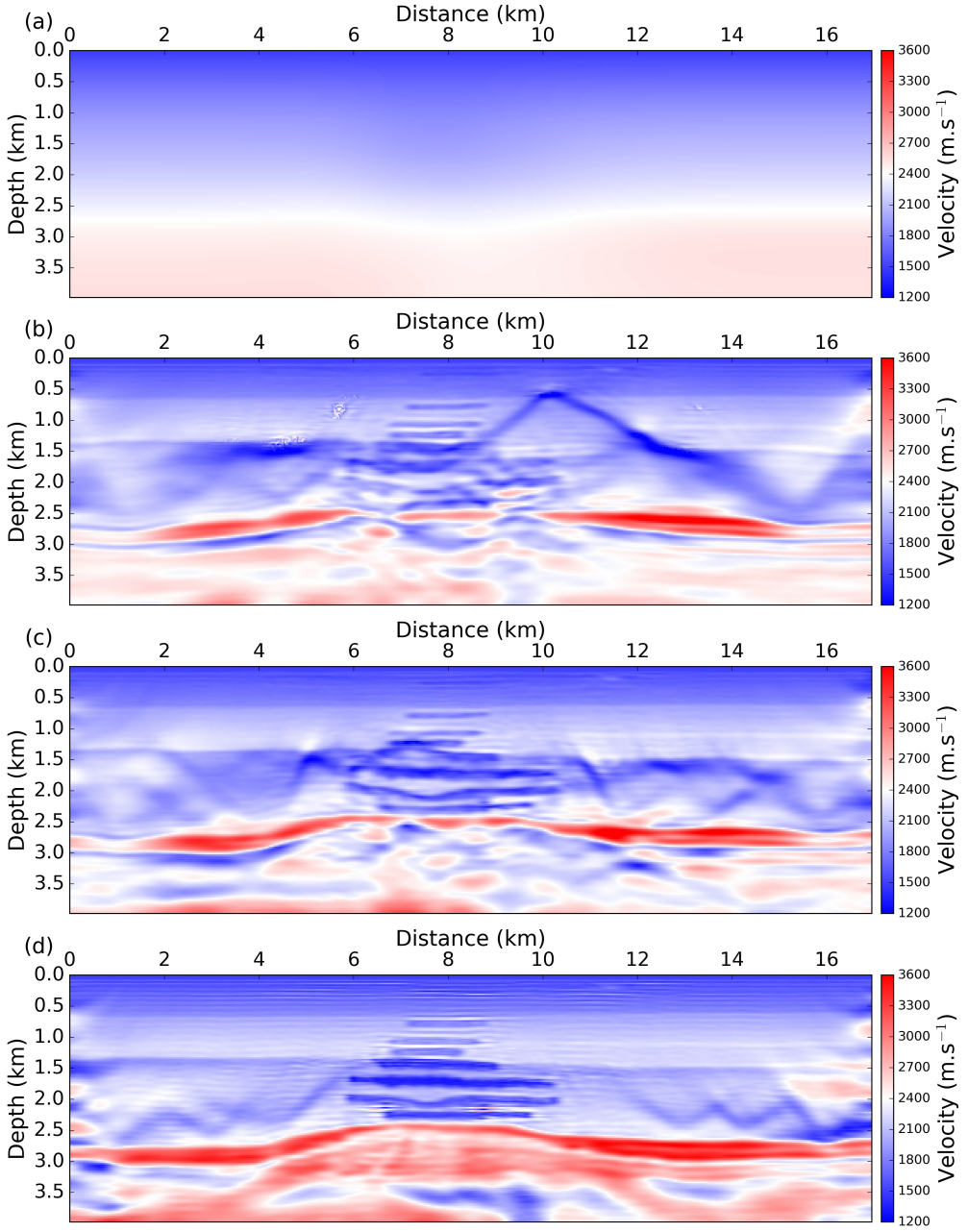

Figure 20: Initial model 4 (a). Corresponding reconstructed models using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function with a hierarchical approach using decreasing values of ΔT: 0.5 s, 0.4 s, 0.2 s, 0.8 s, 0.4 s and 0.2 s.

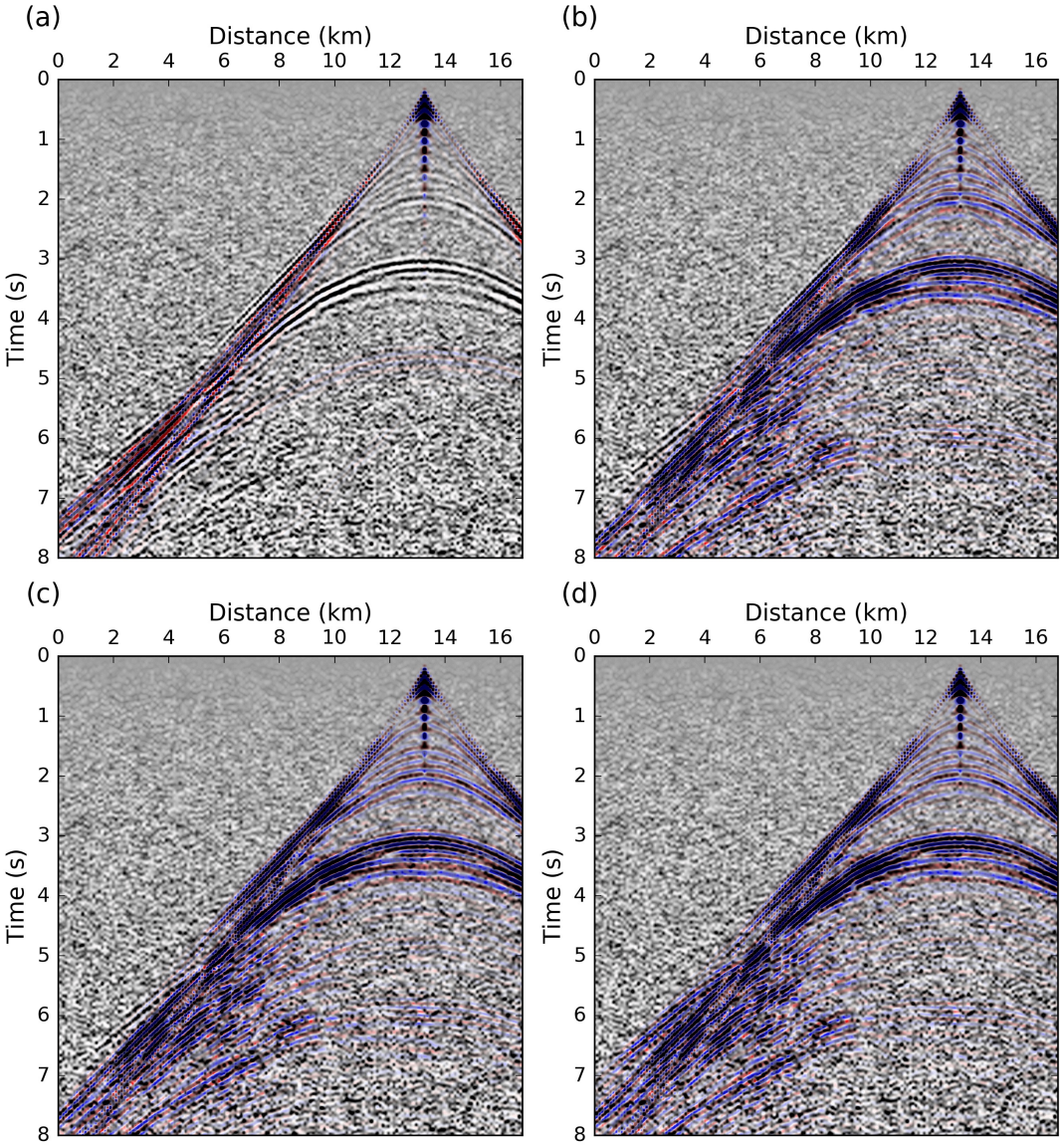

Figure 21: Data fit in the initial model 4 (a), and in the reconstructed models from model 4 using the L^2 misfit function (b), the KR misfit function (c), the graph space OT misfit function within a hierarchical approach with decreasing values of ΔT: 0.5 s, 0.4 s, 0.2 s, 0.8 s, 0.4 s and 0.2 s (d). The real data is represented in black and white, while the synthetic data computed in the corresponding models is represented in red and blue. When the synthetic data predicts correctly the true data, the blue overlays the white events, while the black overlays the red events.

6. Conclusion and Perspectives

We present in this study a graph space OT misfit function for the interpretation of seismic data in the framework of FWI. As initially proposed in Métivier et al. (2018), following the idea by Thorpe et al. (2017), the discrete graph of the observed and synthetic data is compared through OT, which amounts to the comparison of point clouds. This ensures the positivity and mass conservation of the compared quantities, while the resulting misfit function is convex with respect to time and amplitude shifts.

We show how the resulting misfit function can be interpreted as a generalization of a standard L^p distance. The computation of its gradient only requires to access the corresponding adjoint source, which can be computed directly from the solution of the LSAP problem. For practical FWI applications, the computation of the misfit function and the adjoint source amounts to the solution of a sequence of small scale dense LSAP problems, which we solve with the auction algorithm: the computation cost increase compared to L^p distance is limited to few percents for realistic scale applications.

The investigation of the 2D synthetic Valhall case study exhibits the interesting properties of this graph space OT misfit function for FWI. The sensitivity to the accuracy of the initial model is strongly decreased compared to L^p misfit functions or the previously introduced KR approach.

The next step of this investigation will be devoted to its application to a 3D field data from the Valhall field for which we already have a good expertise. This will make possible to assess the interest of this graph space OT approach for 3D imaging of field data.

Methodological questions remain still opened regarding the interpretation of the full shot gather through the graph space approach, instead of considering the trace-by-trace strategy which is developed here. While it seems natural to account for the coherency of the data in the whole 2D shot gather space (as would do the eye of a geophysicist), the size of the corresponding problem would reach $O(10^4)$, far beyond the zone where the auction algorithm can be used with efficiency. Two possibilities might be considered to perform this generalization.

The first consists in localizing the auction algorithm, by restraining the total displacement along the time and receiver axis that might be considered in the solution of the LSAP problem. This is consistent with the scaling strategy we are already using to balance the influence of time and amplitude.

The second consists in considering other solvers, dedicated to large scale instance of OT problems. The semi-discrete approach proposed by (Mérigot, 2011; Lévy, 2015; Kitagawa et al., 2017) appears to us as the most promising approach in this frame.

Finally, we would like to emphasize that the OT graph space strategy developed here might be applied to any PDE-constrained optimization problem where signed data is involved, beyond the application to seismic imaging developed here. Modifying the data misfit measurement to the proposed graph space OT measurement yields the possibility to account for the geometric structure of the data in an automatic way, which can potentially bring substantial benefit in the solution of such inverse problems.

Acknowledgements

The authors would like to address personal thanks to G. Métivier for insightful discussions, help and advice. This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by AKERBP, 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."

Appendix A. Auction algorithm

Input: $\varepsilon > 0$, $\tau > 0$, c_{ij}
Output: γ_{ij}
$\psi_j = 0$ `// Initialization` ε `scaling loop`
while $\varepsilon > \tau$ **do**
  $\gamma_{ij} = 0$ `// Initialization auction loop`
  $U = \{1, \dots, K\}$
  $P(j) = \emptyset, j \in \{1, \dots, K\}$
  while $U \neq \emptyset$ **do**
    `// Bidding phase`
    for $i \in I \subset U$ **do**
      $j_i = \underset{j=1,\dots,K}{\arg\max} \{c_{ij} - \psi_j\}$
      $\varphi_i^1 = c_{ij_i} - \psi_{j_i}$
      $\varphi_i^2 = \underset{j=1,\dots,K, j \neq j_i}{\max} \{c_{ij} - \psi_j\}$
      $b_{ij_i} = \psi_j + \varphi_i^1 - \varphi_i^2 - \varepsilon$
      $P(j_i) = P(j_i) \cup \{i\}$
    end
    `// Assignment phase`
    for $j = 1, \dots, K$ **do**
      if $P(j) \neq \emptyset$ **then**
        $\psi_j = \underset{i \in P(j)}{\max} b_{ij}$
        for $i = 1, \dots, K$ **do**
          if $\gamma_{ij} = 1$ **then**
            $\gamma_{ij} = 0$
            $U = U \cup \{i\}$
          end
        end
        $i_j = \underset{i \in P(j)}{\arg\max}\, b_{ij}$
        $\gamma_{i_j j} = 1$
        $U = U / \{i_j\}$
      end
    end
  end
  $\varepsilon = \varepsilon / 4$
end

Algorithm 1: Standard auction algorithm for the problem (101).

References

Aghamiry, H., Gholami, A., and Operto, S. (2018). Improving full-waveform inversion based on wavefield reconstruction via Bregman iterations. In *Expanded Abstracts, 80th Annual EAGE Meeting (Copenhagen)*.

Akgül, M. (1993). A genuinely polynomial primal simplex algorithm for the assignment problem. *Discrete Applied Mathematics*, 45(2):93–115.

Ambrosio, L. (2003). Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces*, volume 1812 of *Lecture Notes in Mathematics*, pages 1–52. Springer Berlin Heidelberg.

Ambrosio, L., Mainini, E., and Serfaty, S. (2011). Gradient flow of the Chapman Rubinstein Schatzman model for signed vortices. *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis*, 28(2):217–246.

Amestoy, P., Brossier, R., Buttari, A., L'Excellent, J.-Y., Mary, T., Métivier, L., Miniussi, A., and Operto, S. (2016). Fast 3D frequency-domain full waveform inversion with a parallel Block Low-Rank multifrontal direct solver: application to OBC data from the North Sea. *Geophysics*, 81(6):R363 – R383.

Balinski, M. L. (1985). Signature methods for the assignment problem. *Operations Research*, 33(3):527–536.

Benamou, J. D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

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.

Benamou, J. D., Froese, B. D., and Oberman, A. M. (2014). Numerical solution of the Optimal Transportation problem using the Monge-Ampère equation. *Journal of Computational Physics*, 260:107–126.

Bérenger, J.-P. (1994). A perfectly matched layer for absorption of electromagnetic waves. *Journal of Computational Physics*, 114:185–200.

Bertsekas, D. P. (1998). *Network Optimization: Continuous and Discrete Models*. Athena Scientific.

Bertsekas, D. P. and Castanon, D. (1989). The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1):67–96.

Birkhoff, G. (1946). Three observations on linear algebra. *Univ. Nac. Tucumán*, A(5):147–151.

Bozdağ, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N., and Pugmire, D. (2016). Global adjoint tomography: first-generation model. *Geophysical Journal International*, 207(3):1739–1766.

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.

Bretaudeau, F., Brossier, R., Leparoux, D., Abraham, O., and Virieux, J. (2013). 2D elastic full waveform imaging of the near surface: Application to synthetic and a physical modelling data sets. *Near Surface Geophysics*, 11:307–316.

Brossier, R., Operto, S., and Virieux, J. (2009). Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *Geophysics*, 74(6):WCC105–WCC118.

Brossier, R., Operto, S., and Virieux, J. (2010). Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3):R37–R46.

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

Burkard, R., Dell'Amico, M., and Martello, S. (2012). *Assignment Problems*. Society for Industrial and Applied Mathematics.

Cerjan, C., Kosloff, D., Kosloff, R., and Reshef, M. (1985). A nonreflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics*, 50(4):2117–2131.

Chavent, G. (2009). *Nonlinear least squares for inverse problems*. Springer Dordrecht Heidelberg London New York.

Choi, Y. and Shin, C. (2008). Frequency-Domain Elastic Full Waveform Inversion Using the New Pseudo-Hessian Matrix: Experience Of Elastic Marmousi 2 Synthetic Data. *Bulletin of the Seismological Society of America*, 98(5):2402–2415.

Dantzig, G. B. (1991). *Linear Programming and Extensions*. Princeton University Press.

Delon, J. (2006). Movie and video scale-time equalization application to flicker reduction. *IEEE Transactions on Image Processing*, 15(1):241–248.

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

Dominitz, A. and Tannenbaum, A. (2010). Texture mapping via optimal mass transport. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):419–433.

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

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.

Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H. P. (2010). Full waveform tomography for radially anisotropic structure: New insights into present and past states of the Australasian upper mantle. *Earth and Planetary Science Lettters*, 290(3-4):270–280.

Gass, S. I. (1984). *Linear Programming: Methods and Applications (5th Ed.)*. McGraw-Hill, Inc., New York, NY, USA.

Gauthier, O., Virieux, J., and Tarantola, A. (1986). Two-dimensional nonlinear inversion of seismic waveforms: numerical results. *Geophysics*, 51(7):1387–1403.

Groos, L., Schäfer, M., Forbriger, T., and Bohlen, T. (2014). The role of attenuation in 2D full-waveform inversion of shallow-seismic body and Rayleigh waves. *Geophysics*, 79(6):R247–R261.

Hampson, D. (1991). AVO inversion, theory and pratice. *The Leading Edge*, pages 39–42.

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

Karmarkar, N. (1984). A New Polynomial-time Algorithm for Linear Programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pages 302–311, New York, NY, USA. ACM.

Kitagawa, J., Mérigot, Q., and Thibert, B. (2017). Convergence of a Newton algorithm for semi-discrete optimal transport. *ArXiv e-prints*.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(12):83–97.

Lellmann, J., Lorenz, D., Schönlieb, C., and Valkonen, T. (2014). Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859.

Lévy, B. (2015). A numerical algorithm for L2 semi-discrete optimal transport in 3D. *ESAIM: M2AN*, 49(6):1693–1715.

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

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

Makhorin, A. (2017). *GNU Linear Programming Kit Reference Manual for GLPK Version 4.64*. Moscow Aviation Institute, Department for applied informatics.

Megiddo, N. (1989). Pathways to the optimal set in linear programming. In Megiddo, N., editor, *Progress in Mathematical Programming: Interior-Point and Related Methods*, pages 131–158, New York, NY. Springer New York.

Mérigot, Q. (2011). A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592.

Métivier, L., Allain, A., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2018). Optimal transport for mitigating cycle skipping in full waveform inversion: a graph space transform approach. *Geophysics*, 83(5):R515–R540.

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, L., Brossier, R., Labbé, S., Operto, S., and Virieux, J. (2014). A robust absorbing layer for anisotropic seismic wave modeling. *Journal of Computational Physics*, 279:218–240.

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, 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, 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.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Mulder, W. and Plessix, R. E. (2008). Exploring some issues in acoustic full waveform inversion. *Geophysical Prospecting*, 56(6):827–841.

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

Operto, S., Miniussi, A., Brossier, R., Combe, L., Métivier, L., Monteiller, V., Ribodetti, A., and Virieux, J. (2015). Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to Valhall in the visco-acoustic vertical transverse isotropic approximation. *Geophysical Journal International*, 202(2):1362–1391.

Pitié, F., Kokaram, A. C., and Dahyot, R. (2007). Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1):123 – 137. Special issue on color image processing.

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.

Plessix, R. E. and Perkins, C. (2010). Full waveform inversion of a deep water ocean bottom seismometer dataset. *First Break*, 28:71–78.

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.

Qiu, L., Ramos-Martìnez, J., Valenciano, A., Yang, Y., and Engquist, B. (2017). Full-waveform inversion with an exponentially encoded optimal-transport norm. In *SEG Technical Program Expanded Abstracts 2017*, pages 1286–1290.

Rabin, J., Peyré, G., and Cohen, L. D. (2010). Geodesic Shape Retrieval via Optimal Mass Transport. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *Computer Vision – ECCV 2010*, pages 771–784, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein Barycenter and Its Application to Texture Mixing. In Bruckstein, A. M., ter Haar Romeny, B. M., Bronstein, A. M., and Bronstein, M. M., editors, *Scale Space and Variational Methods in Computer Vision*, pages 435–446, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

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.

Schäfer, M., Groos, L., Forbriger, T., and Bohlen, T. (2013). 2D full waveform inversion of recorded shallow seismic Rayleigh waves on a significantly 2D structure. In *Proceedings of 19th European Meeting of Environmental and Engineering Geophysics, Expanded Abstracts, Bochum, Germany*.

Schmitzer, B. (2016). A Sparse Multiscale Algorithm for Dense Optimal Transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259.

Shin, C., Jang, S., and Min, D. J. (2001). Improved amplitude preservation for prestack depth migration by inverse scattering theory. *Geophysical Prospecting*, 49:592–606.

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.

Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., and Kommedal, J. H. (2010). Full waveform inversion: the next leap forward in imaging at Valhall. *First Break*, 28:65–70.

Stopin, A., Plessix, R.-E., and Al Abri, S. (2014). Multiparameter waveform inversion of a large wide-azimuth low-frequency land data set in Oman. *Geophysics*, 79(3):WA69–WA77.

Symes, W. (2015). Algorithmic aspects of extended waveform inversion. In *77th EAGE Conference and Exhibition 2017-Workshops*.

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

Tape, C., Liu, Q., Maggi, A., and Tromp, J. (2010). Seismic tomography of the southern California crust based on spectral-element and adjoint methods. *Geophysical Journal International*, 180:433–462.

Thorpe, M., Park, S., Kolouri, S., Rohde, G. K., and Slepcev, D. (2017). A transportation L^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59:187–210.

van Leeuwen, T. and Herrmann, F. (2016). A penalty method for PDE-constrained optimization in inverse problems. *Inverse Problems*, 32(1):1–26.

van Leeuwen, T. and Mulder, W. A. (2010). A correlation-based misfit criterion for wave-equation traveltime tomography. *Geophysical Journal International*, 182(3):1383–1394.

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

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*, pages R1–1–R1–40. Society of Exploration Geophysics.

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

Wang, C., Yingst, D., Farmer, P., and Leveille, J. (2016). *Full-waveform inversion with the reconstructed wavefield method*, pages 1237–1241.

Wang, Y. and Rao, Y. (2009). Reflection seismic waveform tomography. *Journal of Geophysical Research*, 114(B3):1978–2012.

Warner, M. and Guasch, L. (2016). Adaptive waveform inversion: Theory. *Geophysics*, 81(6):R429–R445.

Yang, P., Brossier, R., Métivier, L., and Virieux, J. (2016a). A review on the systematic formulation of 3D multiparameter full waveform inversion in viscoelastic medium. *Geophysical Journal International*, 207(1):129–149.

Yang, P., Brossier, R., Métivier, L., and Virieux, J. (2016b). Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method. *Geophysics*, 81(6):R349–R362.

Yang, P., Brossier, R., Métivier, L., Virieux, J., and Zhou, W. (2018a). A Time-Domain Preconditioned Truncated Newton Approach to Multiparameter Visco-acoustic Full Waveform Inversion. *SIAM Journal on Scientific Computing*, 40(4):B1101–B1130.

Yang, Y. and Engquist, B. (2018). Analysis of optimal transport and related misfit functions in full-waveform inversion. *GEOPHYSICS*, 83(1):A7–A12.

Yang, Y., Engquist, B., Sun, J., and Hamfeldt, B. F. (2018b). Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–R62.