Th SRS2 01

Overcoming Cycle Skipping in FWI - An Optimal Transport Approach

L. Metivier* (Univ. Grenoble Alpes/CNRS), R. Brossier (Univ. Grenoble Alpes), Q. Mérigot (Univ. Paris Dauphine/CNRS), E. Oudet (Univ. Grenoble Alpes) & J. Virieux (Univ. Grenoble Alpes)

SUMMARY

Conventional full waveform inversion based on the L2 norm is known to suffer from cycle skipping. Numerous techniques have been promoted to face this long term issue, relying on modifications of the function which measures the discrepancy between observed and predicted data. In this study, we propose to measure this discrepancy through the solution of an optimal transport problem. Instead of proceeding to a sample-by-sample comparison of the seismic signal, this offers the possibility to perform a global comparison of the data, taking into account the spatial and time coherency of the predicted and observed shot gathers. The optimal transport problem is reformulated as a non-smooth convex optimization problem, which can be solved efficiently through proximal splitting techniques. As for other modifications of the misfit function, the optimal transport distance is integrated naturally in the FWI workflow through a modification of the adjoint source term. This modified adjoint source is simply the solution of the optimal transport problem. A 2D time-domain acoustic application on the Marmousi 2 model is presented. We show how the sensitivity to the accuracy of the initial model can be reduced using the optimal transport distance.

Introduction

Full waveform inversion is a seismic imaging method designed for the computation of high resolution estimations of the subsurface mechanical properties (P and S-wave velocities, density, attenuation, anisotropy parameters). This technique is conventionally based on the minimization of a misfit function measuring the L^2 distance between predicted and observed data. When applied to the reconstruction of velocities, FWI is known to suffer from the non-convexity of this misfit function. The local minima correspond to subsurface models allowing to match the observed data up to one or several phase shifts (cycle skipping problem). For realistic size FWI applications, only local descent minimization techniques are computationally affordable. Converging towards the global minimum thus requires to start from an initial model accurate enough to match the observed data within half a phase.

In order to further improve the reliability of FWI in case of cycle skipping, two main strategies have been developed: image domain techniques, based on focusing the energy in the migrated domain with the purpose to reconstruct smooth background models (Symes, 2008); data-domain techniques based on the modification of the misfit measurement. Cross-correlation (Luo and Schuster, 1991) and time warping techniques (Ma and Hale, 2013) focus for instance on time delays between arrivals, which yields more convex misfit functions, at the cost of a resolution loss. Further developments in this direction have led to deconvolution based approaches (Luo and Sava, 2011), using Wiener filter to match observed and predicted data. Recent applications have demonstrated the capability of this method to mitigate cycle skipping in 2D (Guasch and Warner, 2014).

In this study, an alternative data-domain technique is proposed. The discrepancy between seismograms is evaluated using a distance related to the optimal transport theory. This theory originates from the work of the French engineer Gaspard Monge, in an attempt to devise the best strategy to move sand to a building site. In short, the optimal transport distance, also know as Earth Mover's Distance, or Wasserstein distance, considers all the transformations allowing to map the predicted data to the observed data. These transformations (mappings) are seen as an ensemble of mass transportation between two points. A cost is associated to each of these transformations: this cost is related to the amount of mass transported between two points and the distance between these two points. The Wasserstein distance corresponds to the minimal cost among all the possible transformations.

The standard L^2 distance proceeds in a sample-by-sample comparison of the seismograms, discarding all the information related to the time and space coherency of the seismic gathers. Conversely, the Wasserstein distance relies on a global comparison of the shot-gather images, which offers the possibility of taking into account this information. This is the main motivation for using this distance in the framework of FWI. We present how this distance can be used to compare seismograms. An efficient numerical strategy based on proximal splitting techniques and a multi-grid algorithm is designed to compute this distance on realistic size seismograms. A synthetic case study in a cross-hole tomography configuration is first presented to investigate the convexity of the misfit function, under the assumption of an homogeneous velocity model. Results obtained on the Marmousi 2 model are then presented, emphasizing the better properties of the optimal transport distance compared to the L^2 distance in terms of cycle skipping issues.

Theory

The Wasserstein distance has been originally designed for the comparison of probability measures: this implies that the signals which are compared are positive and that the total mass (energy of the signal) is conserved. These two assumptions are obviously not satisfied when comparing seismic signals in FWI: the data are oscillating between positive and negative values, and there is no reason to assume that the total energy of the predicted data is the same as the total energy of the observed data.

The proposed strategy consists in using a distance allowing for energy non-conservation and non-positive signals while still preserving the properties of the Wasserstein distance. This distance is based on the Kantorovich-Rubinstein norm $\|.\|_{KR}$ (Lellmann et al., 2014). We consider two predicted and observed shot-gathers, each constituted of N_r traces containing N_t discrete time samples, denoted by d_{pred} and d_{obs}. The distance between d_{pred} and d_{obs} is computed as

$$\begin{aligned} &\left\|d_{pred}-d_{obs}\right\|_{KR}=\max_{\varphi}\sum_{i=1}^{N_r}\sum_{n=1}^{N_t}\varphi_i^n\left((d_{pred})_i^n-(d_{obs})_i^n\right)\equiv\left\langle\varphi,d_{pred}-d_{obs}\right\rangle \\ &s.t.\ \forall(i,j,n,m),\ \ 1\leq(i,j)\leq N_r,\ \ 1\leq(n,m)\leq N_t,\ \ |\varphi_i^n-\varphi_j^m|<|x_i-x_j|+|t_n-t_m|,\ \ |\varphi_i^n|<1. \end{aligned} \tag{1}$$

The first set of constraints ensure that the function φ is 1-Lipschitz. Relaxing the second set of constraints (bound constraints) brings back to the standard dual formulation of the Wasserstein distance (Villani, 2003). Using a property of the ℓ_1 distance, the 1-Lipschitz property can be imposed using only local constraints (Métivier et al., 2016), yielding the simplified linear programming problem

$$\begin{aligned} &\left\|d_{pred}-d_{obs}\right\|_{KR}=\max_{\varphi}\left\langle\varphi,d_{pred}-d_{obs}\right\rangle \\ &s.t.\ \forall(i,n),\ \ 1\leq i\leq N_r,\ \ 1\leq n\leq N_t,\ \ |\varphi_{i+1}^n-\varphi_i^n|<\Delta x,\ \ |\varphi_i^{n+1}-\varphi_i^n|<\Delta t,\ \ |\varphi_i^n|<1, \end{aligned} \tag{2}$$

where Δx and Δt are the discretization steps in the receiver and time dimension respectively. This problem is rewritten under the form of the convex non-smooth optimization problem

$$\max_{\varphi}\left\langle\varphi,d_{pred}-d_{obs}\right\rangle+i_K(A\varphi), \tag{3}$$

where K is the unit hypercube of $\mathbb{R}^{3N}$, $K=\left\{x\in\mathbb{R}^{3N},\ \ |x_i|<1\right\}$, N is the total number of discrete points $N=N_r\times N_t$, i_K is the indicator function of the ensemble K

$$i_K(x)=\left|\begin{array}{lll}0 & \text{if} & x\in K\\ +\infty & \text{if} & x\notin K,\end{array}\right. \tag{4}$$

and A is the matrix encoding the linear constraints of (2). This matrix can be written as $A=[D_x\quad D_t\quad I]^T$ where D_x and D_t are the discrete backward difference operators in x and t direction respectively, while I is the identity matrix of size N. The problem (3) is solved efficiently using the proximal splitting method named SDMM (Combettes and Pesquet, 2011). Each iteration of the SDMM algorithm requires the application of the proximity operators of $h(\varphi)\equiv\left\langle\varphi,d_{pred}-d_{obs}\right\rangle$ and i_K which have closed-form. The proximity operator of $h(\varphi)$ is equal to $\varphi-d_{cal}+d_{obs}$. The proximity operator of i_K is simply the projection onto the hypercube K. The most time consuming part of the SDMM algorithm is the solution of a linear system involving the matrix A^TA at each iteration. However, from the expression of A, it is straightforward to see that

$$A^TA=D_x^TD_x+D_t^TD_t+I, \tag{5}$$

which is nothing else than the discretized form of the 2D Laplacian operator with Neumann boundary conditions (plus a constant diagonal term). The linear system (5) is thus a discretized Poisson equation which can be solved in linear complexity through multigrid algorithms (Adams, 1989).

Considering a data-set constituted of S common shot gathers, the FWI problem consists in minimizing the misfit function

$$f_*(m)=\sum_{s=1}^{S}\|d_{pred,s}(m)-d_{obs,s}\|_*, \tag{6}$$

where the index of the shot-gather is denoted by s and $*$ corresponds either to the conventional L^2 norm or to the KR norm. In both cases, the gradient of the misfit function $f_*(m)$ can be computed through the adjoint-state method (Plessix, 2006). The only difference relies on the definition of the source of the adjoint wavefields. While in the L^2 case, this source is equal to the residuals $d_{pred,s}-d_{obs,s}$, using the optimal transport distance, this source term is given by

$$\frac{\partial\max_{\varphi_s}\left\langle\varphi_s,d_{pred,s}(m)-d_{obs,s}\right\rangle}{\partial d_{pred,s}}=\overline{\varphi}_s(m), \tag{7}$$

where $\overline{\varphi}_s(m)$ is the solution of the optimal transport problem (3).

Numerical examples

We start with a simple cross-hole synthetic experiment. A source is located at 2.5 km depth in a first borehole, and an array of 196 receivers equally spaced each 25 m is located in a borehole 2.5 km apart the first borehole. A Ricker source centered on 5 Hz is used to generate a synthetic shot gather in a reference homogeneous background velocity $v_P^*=2000$ m.s^{-1}. Using the assumption that v_P is homogeneous, the

functions f_{L^2} and f_{KR} become functions of one variable, and we compare their profiles in Figure 1a. They both reach the global minimum at $v = 2000$ m.s^{-1}. The function f_{L^2} presents two secondary minima at $v_P = 1780$ m.s^{-1} and $v_P = 2300$ m.s^{-1} (cycle skipping). The secondary minima still exist for f_{KR}, however, they are lifted up, rendering the misfit function closer from a convex function. The valley of attraction remains also as sharp as for f_{L^2}, which ensures that the "resolution power" of the method is unchanged, contrary to wave equation tomography strategies.

Results obtained on the Marmousi 2 model in the acoustic approximation with constant density are presented in Figure 1b. A surface acquisition with 168 receivers every 100 m is used. A Ricker source centered on 5 Hz is used, its frequency content below 2.5 Hz being removed using a minimum phase Butterworth filter. Two initial models are used, obtained after smoothing the exact model with a Gaussian filter with a correlation length equal to 0.25 km and 2 km respectively. The first model is very close from the exact model, with only smoother interfaces. The second model presents almost only vertical variations, and underestimates the increase of the velocity in depth.

Starting from the first initial model, FWI based on both f_{L^2} and f_{KR} yields satisfactory estimations. A difference can be noted regarding the reconstruction of the low velocity zone near $x = 11$ km and $z = 2.5$ km. A high velocity artifact can be seen in this zone in the estimation obtained with f_{L^2}. This is not the case in the estimation obtained with f_{KR}. Starting from the second initial model, FWI based on f_{L^2} is unable to provide a satisfactory P-wave velocity estimation due to sever cycle skipping problems. In comparison, the P-wave velocity estimation obtained using f_{KR} is significantly closer from the exact model (Fig. 1b). Low velocity artifacts, typical of cycle skipping, can still be seen in depth, below 3 km. Low wavenumber artifacts are also visible on the left part of the model ($x < 1$ km). However, in the central part, the P-wave velocity model is correctly recovered, even starting from this crude approximation.

In Figure 2a, the L^2 residuals in the two initial models are compared with their optimal transport counterpart ($\overline{\varphi}(m)$ in (7)). The optimal transport residuals are smoother than the L^2 residuals, with a lower frequency content. An emphasis of particular seismic events in the optimal transport residuals is also noticeable, compared to the L^2 residuals. This is mainly observable for the reflections around 3 s and 8 km offset, for both initial model. The optimal transport thus weights differently the uninterpreted part of the seismograms. The impact of this modification is presented in Figure 2b. The steepest descent updates (opposite of the gradient) associated with f_{L^2} and f_{KR}, computed in the second initial model, are compared in Figure 2b. Cycle skipping can be detected in the L^2 gradient through the strong shallow low velocity updates, in a zone where the velocity should be increased. The optimal transport distance efficiently mitigates these strong artifacts. The energy in depth is also better balanced: the main interfaces constituting the Marmousi model appear in this velocity update.

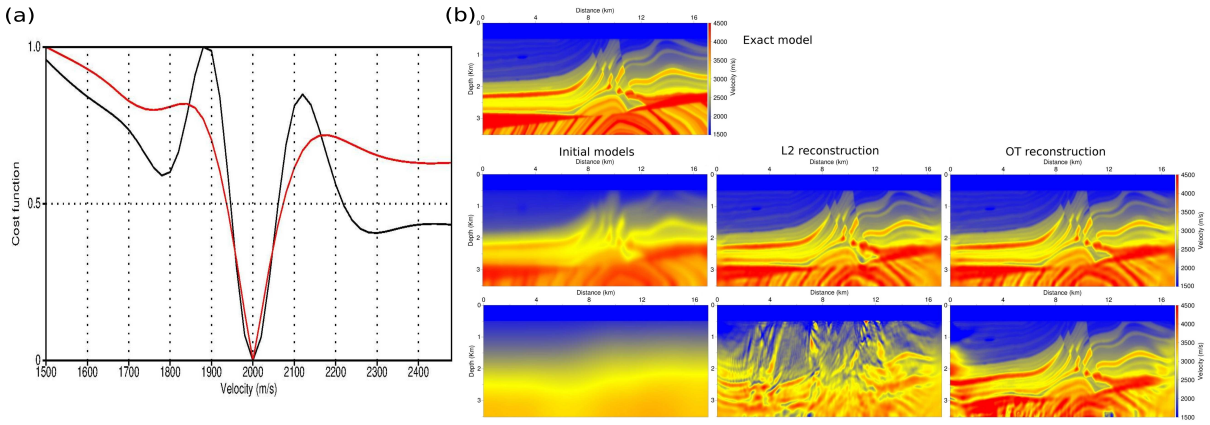

Figure 1 *(a) f_{L^2} profile (black) and f_{KR} profile (red) for the cross-hole tomography experiment. (b) Marmousi 2 case study. Exact and initial models (left column), f_{L^2} results (center column), f_{KR} results (right column).*

Conclusion

A misfit function based on optimal transport is proposed to overcome cycle skipping limitations due to the non convexity of the L^2 norm. The resulting misfit function seems to be more convex and the

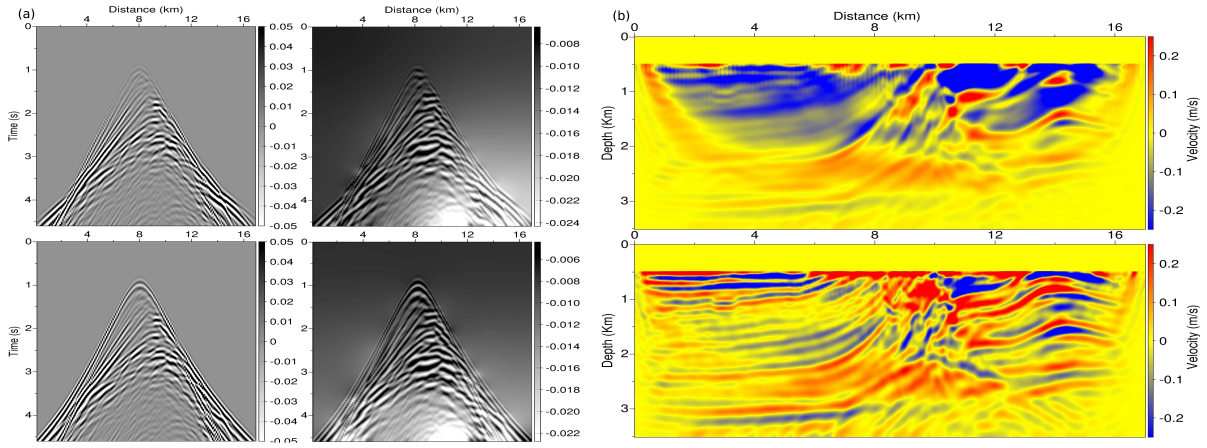

Figure 2 *(a) L^2 residuals (left column) and optimal transport residuals (right column) in the two initial models. (b) First velocity update starting from the second initial model using f_{L^2} (top) and f_{KR} (bottom).*

corresponding FWI strategy less prone to cycle skipping. These first results on synthetic examples provide encouraging perspectives. Results obtained using the same strategy on the BP 2004 salt model and on the Chevron 2014 benchmark dataset are presented in Métivier et al. (2016). The interest of the method needs now to be assessed in more realistic configurations, involving 3D configurations. A more systematic study of the sensitivity of the method with respect to the acquisition design, the level of noise, and the numerical strategy used to solve the optimal transport problem will be performed in the near future.

Acknowledgements

This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by CGG, CHEVRON, EXXON-MOBIL, JGI, PETROBRAS, SCHLUMBERGER, SHELL, SINOPEC, STATOIL, TOTAL and WOODSIDE. This study was granted access to the HPC resources of CIMENT infrastructure (https://ciment.ujf-grenoble.fr) and CINES/IDRIS under the allocation 046091 made by GENCI.

References

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

Combettes, P.L. and Pesquet, J.C. [2011] Proximal Splitting Methods in Signal Processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R. and Wolkowicz, H. (Eds.) *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, *Springer Optimization and Its Applications*, 49, Springer New York, 185–212.

Guasch, L. and Warner, M. [2014] Adaptive Waveform Inversion - FWI Without Cycle Skipping - Applications. In: *76th EAGE Conference and Exhibition 2014*.

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

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.

Ma, Y. and Hale, D. [2013] Wave-equation reflection traveltime inversion with dynamic warping and full waveform inversion. *Geophysics*, **78**(6), R223–R233.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. and Virieux, J. [2016] Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Journal International*, in press.

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.

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

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