New insights on the graph space optimal transport distance for full waveform inversion

L. Métivier[1,2]*, R. Brossier*[2]
[1]*Univ. Grenoble Alpes, CNRS, LJK, F-38058 Grenoble, France*
[2]*Univ. Grenoble Alpes, ISTerre, F-38058 Grenoble, France*

SUMMARY

Non-convexity issues in full waveform inversion is a topic still deserving significant research efforts. One direction relies on modifying the function measuring the distance between observed and synthetic data on which is based the full waveform inversion process. Recently, optimal transport distances have been considered to play this role. As optimal transport theory has been developed for the comparison of positive functions, adaptation needs to be brought to apply it to the comparison of seismic data which are oscillatory. Among different propositions, the graph space optimal transport distance consists in considering each seismic trace as a point cloud in a time/amplitude two-dimensional space. The method has shown interesting properties in application both to synthetic and three-dimensional field data. In this abstract, we present new insights on this misfit function. We first provide a theoretical comparison with the dynamic time-warping approach. We propose a novel formulation of the graph space optimal transport problem making its application more flexible. We demonstrate the simple form of the second-order derivatives of the corresponding misfit function, making it possible to use standard preconditioning method such as pseudo-Hessian which is illustrated on a synthetic experiment with the Marmousi model.

INTRODUCTION

Full waveform inversion is a high resolution seismic imaging method formulated as a partial-differential-equations (PDE) constrained optimization problem. The distance between observed data and synthetic data computed through the solution of PDE representing the wave propagation is minimized over a space of parameters describing the subsurface, *i.e.* wave velocities. This distance is by default chosen as the least-squares distance. This is problematic as the resulting misfit function is non-convex Jannane et al. (1989). Indeed, the size of the associated discrete problem requires the use of local optimization solvers for its solution. This leads to the potential convergence towards non-informative local minima, an issue often referred to as cycle-skipping in the FWI community.

To overcome this issue, multi-scale hierarchical workflow are widely used in practice, together with the use of accurate initial model building tools such as stereotomography (Billette and Lambaré, 1998). Theses process are not always successful, for instance because of the lack of sufficiently good quality low frequency data at the exploration scale. Even when consistent results are obtained, they often rely on complicated multi-steps workflow requiring strong human expertise. This in turns increase the uncertainty attached to these models: how a change in the workflow would affect the final result?

This calls for more robust full waveform inversion methods. Two main directions are currently investigated: the use of extended model strategies or misfit function modifications. In both cases, the convexity of the optimization problem is the motivation. In the frame of misfit function modifications, the use of optimal transport distances has attracted attention recently (Engquist and Froese, 2014). Optimal transport distances are convex with respect to translation and dilation in the functions they compare. Applied in the frame of full waveform inversion, convexity with respect to time shifts is a good proxy towards convexity with respect to subsurface velocities.

However, optimal transport distances are defined for the comparison of probability distributions, and applying them to seismic data requires care. While it is possible to modify the data through non-linear transform and normalization techniques prior to the comparison with optimal transport (Yang and Engquist, 2018), we have been interested in alternative ways of applying OT to seismic data. Our motivation is that such non-linear transforms alter the shape of the data, which might result in uncontrolled sensitivity of the misfit function with respect to specific seismic events in the data. When considering noisy field data, this uncontrolled behavior can be problematic.

Aside using a specific instance of optimal transport distance (namely the 1-Wasserstein distance or the Kantorovich-Rubinstein norm), which loses the convexity with respect to time shifts but makes it possible to perform global comparison of the data in a multi-dimensional space, taking into account the lateral coherency of seismic gathers in the time/position plane (Métivier et al., 2016), we have proposed a lift toward the graph of the data. In this frame, each seismic trace is interpreted, after discretization, as a point cloud in a 2D time/amplitude space. This point cloud is mathematically represented as a sum of Dirac measures, therefore a positive function. Optimal transport is thus applied to the comparison of point clouds associated with synthetic and observed traces. We have named this distance: graph space optimal transport distance (Métivier et al., 2019).

This distance has now been applied successfully to 2D synthetic data from Marmousi, BP2004, Valhall models, to the 2D Chevron model from 2014 in a reflection waveform inversion frame Provenzano et al. (2020) as well as to 2D field data from the Nankai trough (Górszczyk et al., 2020) and 3D field data from Valhall (Pladys et al., 2020). The aim of this study is to provide novel elements of analysis related to this misfit function. We first draw a comparison with dynamic time-warping (DTW) approach from (Ma and Hale, 2013), demonstrating how the two methods are intimately related. Then, we discuss how the scaling between time and amplitude axis can be designed to compare the point clouds associated with the discrete graph of the seismic traces. We present a novel scaling strategy making the application of GSOT more flexible and unlocking its application toward the comparison of full multi-dimensional seismic gathers. Finally, we show how the theorem behind the expression of the gradient of the GSOT misfit function can be used to obtain a simple expression of the Hessian operator. This simple expression makes it possible to design preconditioners for the GSOT misfit function based on approximations of the conventional Gauss-Newton operator. We illustrate the latter aspect using the Marmousi model.

GRAPH SPACE OPTIMAL TRANSPORT DISTANCE FORMULATION AND LINK WITH DYNAMIC TIME WARPING

Consider one seismic trace $d(t)$. We assume it is regularly sampled with N discretization points and a time discretization step dt. The discrete graph of $d(t)$ is the ensemble of points of $\mathbb{R}^2$

$$(t_i, d(t_i)), \;\; i = 0, \ldots, N, \tag{1}$$

with $t_i = i \times dt$.

For two traces $d_1(t)$ and $d_2(t)$, the GSOT misfit function $g(d_1, d_2)$ corresponds to the 2-Wasserstein distance between the point clouds of their discrete graphs. The 2-Wasserstein distance between these point clouds corresponds to the solution of the following optimal assignment problem

$$g(d_1, d_2) = \min_{\sigma \in S(N)} \sum_{i=0}^{N} c_{i,\sigma(i)}, \tag{2}$$

where $S(N)$ is the ensemble of permutation of $\{1, \ldots, N\}$ and c_{ij} corresponds to the Euclidean distance between two points

$$c_{ij} = |t_i - t_j|^2 + \eta |d_1(t_i) - d_2(t_j)|^2. \tag{3}$$

In 3, η is a dimensioning parameter which controls the convexity of the GSOT misfit function with respect to time shifts.

10.1190/segam2021-3583678.1

The GSOT FWI misfit function is built as a summation over each trace, namely each source/receiver couple, such that the corresponding FWI problem is

$$\min_m f(m) = \sum_{s,r} \alpha_{s,r} g(d_{cal_s,r}[m], d_{obs,s,r}), \tag{4}$$

where

$$d_{cal,s,r}[m] = R_r u_s[m], \;\; A(m)u_s = b_s, \tag{5}$$

with R_r an operator extracting the values of the wavefield $u_s[m]$ at receiver position r, $A(m)$ a partial differential operator representing the wave propagation within the subsurface, m a parameter of this PDE *i.e.* seismic wave velocity, density, attenuation, and b_s a source term associated with source s. The parameters $\alpha_{s,r}$ are introduced to restore the AVO information which might be lost in the dimensioning through η. This issue is discussed in the next Section in details.

Now, we introduce the ensemble D_S of time shift functions $h(t)$ related to an assignment $\sigma \in S(N)$ such that

$$D_S = \left\{ h(t), \;\; \exists \sigma \in S(N), \;\; h(t_i) = t_i - t_{\sigma(i)}, \;\; i = 1, \ldots, N. \right\} \tag{6}$$

Using D_S, one can rewrite 2 using continuous notations as

$$g(d_1, d_2) = \min_{h \in D_S} \int_0^T \eta |d_1(t) - d_2(t - h(t))|^2 dt + \|h\|^2, \tag{7}$$

where $\|.\|$ is the least-squares norm for real-valued functions defined on $[0, T]$

$$\|h\|^2 = \int_0^T |h(t)|^2 dt. \tag{8}$$

From 7 one can see a close connection with DTW (Ma and Hale, 2013). In this approach, a time-dependent time shift function $h(t)$ is computed to define the distance between two traces, such that

$$\widetilde{g}(d_1, d_2) = \min_{h \in C} \int_0^T |d_1(t) - d_2(t - h(t))|^2 dt, \tag{9}$$

with C an ensemble of constraints related to $h(t)$, namely bound constraints and smoothness constraints to stabilize the solution.

The main difference between GSOT and DTW is thus the space to which the optimal time shift function $h(t)$ belongs and the presence of a regularization term in GSOT. In the latter approach, the time shift function is parameterized in a particular way depending on a permutation $\sigma \in S(N)$, and the least-squares norm of the time shift $h(t)$ is penalized so as to regularize the misfit function. In DTW, there is no such penalization term, but the solution space is different and does not rely on a specific permutation. For instance, no crossings between events are allowed in DTW. The same number of events is expected and a local time shift for each event is computed. This is not the case with GSOT. In the latter approach, the solution space D_S might appear peculiar, however it is inherited from the OT theory. This guarantees existence and uniqueness of the solution, the distance properties for the GSOT function (identity of indiscernibles, symmetry and triangular inequality), and a tractable computation through linear assignment solvers such as the auction algorithm of Bertsekas and Castanon (1989). This is not the case for DTW to the best of our knowledge.

WEIGHTING STRATEGY

In 3, the scaling parameter η plays a crucial role. It controls the behavior of the GSOT misfit function by weighting the cost of assigning points of the graph of d_1 and d_2 along the amplitude axis. If η is chosen to be "small", the assignment is preferably done along the amplitude axis, and the GSOT misfit function boils down to the conventional least-squares distance. On the other hand, if η is "large", the assignment is preferably done along the time axis, and the GSOT misfit function becomes sensitive to time shifts.

More precisely, a practical choice for η is, for a trace s, r,

$$\eta = \eta_{s,r} = \frac{\tau^2}{A_{s,r}^2}, \tag{10}$$

where τ is a maximum expected time shift and $A_{s,r}$ is an amplitude normalization parameter, for instance the maximum peak amplitude difference between $d_{cal,s,r}$ and $d_{obs,s,r}$. Following this definition, a point $\left(t, d_{cal,s,r}(t)\right)$ such that its amplitude difference with $d_{obs,s,r}(t)$ is equal to $A_{s,r}$ can be assigned with a point of same amplitude shifted by τ' such that $|\tau'| \leq \tau$. If $|\tau'| > \tau$ then it will be assigned with $\left(t, d_{obs,s,r}(t)\right)$. In particular, we see that for a specific phase with an amplitude difference A smaller than $A_{s,r}$, the "effective" maximum time shift allowed for this phase is smaller than τ. More precisely, it becomes equal to $\tau \times A/A_{s,r}$. Therefore the smallest the amplitude difference, is the smallest the maximum "effective" time shift becomes. This interpretation is illustrated in Figure 1.

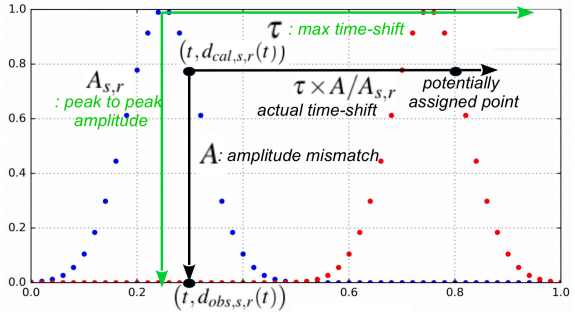

Figure 1: Scaling interpretation for the comparison of the point clouds associated with two shifted Gaussian functions. The maximum peak amplitude difference is $A_{s,r}$ and the actual amplitude difference for the considered point $\left(t, d_{cal,s,r}(t)\right)$ is A. The maximum possible time shift for this point is thus equal to $\tau \times A/A_{s,r}$. In the illustration, a point from the point clouds $\left(t, d_{obs,s,r}(t)\right)$ lies within this distance and therefore could be assigned to $\left(t, d_{cal,s,r}(t)\right)$.

This scaling has another important effect. Applied to each trace s, r independently, it acts as a trace-by-trace normalization which discards the AVO information. This is visible for instance in the adjoint source expression, on which is based the FWI gradient computation (Plessix, 2006). Denoting $\sigma^*_{s,r}$ the solution of the assignment problem for the trace s, r, we have shown in (Métivier et al., 2019) that the adjoint source for this trace is given by

$$\frac{\partial g\left(d_{cal,s,r}, d_{obs,s,r}\right)}{\partial d_{cal,s,r}} = \frac{2\tau^2}{A_{s,r}^2}\left(d_{cal,s,r}(t) - d^{\sigma^*}_{obs,s,r}(t)\right), \tag{11}$$

where $d^{\sigma^*}_{obs,s,r}(t_i) = d_{obs,s,r}(t_{\sigma^*(i)})$. The adjoint source thus corresponds to the difference between calculated and observed data, for samples connected through the assignment σ^*, normalized by the trace-dependent factor $2\tau^2/A^2_{s,r}$.

To re-inject the AVO trend in the misfit function, we have proposed to introduce additional scaling parameters $\alpha_{s,r}$ to re-weight the contribution of each trace in the assembly of the misfit function. In this study, we propose an alternative scaling strategy which naturally preserves the AVO information and the trace amplitude. Following this strategy, the distance between two points c_{ij} becomes

$$c_{ij} = \frac{A^2_{s,r}}{\tau^2}|t_i - t_j|^2 + |d_1(t_i) - d_2(t_j)|^2. \tag{12}$$

Let us analyze this simple change. First, the solution σ^* of the assignment problem 2 remains the same. A point $(t, d_{cal,s,r}(t))$ is equally distant from a point $(t \pm \tau, d_{cal,s,r}(t))$ or a point $(t, d_{obs,s,r}(t))$ with $A_{s,r}$ still the amplitude difference $A_{s,r} = |d_{cal,s,r}(t) - d_{obs,s,r}(t)|$. The difference is now that the dimensioning is done taking amplitude as the reference scale, while the previous dimensioning was done taking time as the reference scale. The interest for this change is that the relative trace-by-trace amplitude is now preserved. Indeed, the corresponding adjoint source is now

$$\frac{\partial g\left(d_{cal,s,r}, d_{obs,s,r}\right)}{\partial d_{cal,s,r}} = 2\left(d_{cal,s,r}(t) - d^{\sigma^*}_{obs,s,r}(t)\right), \tag{13}$$

from which the previous amplitude scaling is removed. The re-scaling parameters $\alpha_{r,s}$ are thus now useless.

Using this strategy, it is now possible to design more specific scaling and move towards anisotropic metrics c_{ij} to measure the distance

10.1190/segam2021-3583678.1

between points. For instance, one could consider a time-dependent scaling, such as

$$c_{ij} = \frac{A^2_{s,r}(t_i)}{\tau^2}|t_i - t_j|^2 + |d_1(t_i) - d_2(t_j)|^2, \quad (14)$$

to adapt the "effective" maximum time shifts to phases of specific amplitudes within the trace. Alternatively or additionally, one could adapt the expected time shift τ depending on time. Using the original formulation 3 would have induced a time-dependent normalization of the adjoint source *which could not have been compensated through the definition of scaling parameters $\alpha_{r,s}$ in the misfit function*.

This property also opens the way to more easily consider GSOT for the comparison of full seismic gathers. Consider for instance a shot gather $d(t,x)$. Assuming N_r receivers and a time discretization leading to N time samples per trace, its discrete graph is the cloud of $N \times N_r$ points $\big(t_i, x_j, d(t_i,x_j)\big) \in \mathbb{R}^3$. The GSOT distance between two such shot gathers $d_1(x,t)$ and $d_2(x,t)$ can be formulated as

$$\widehat{g}(d_1,d_2) = \min_{\sigma \in S(N\times N_r)} \sum_{I=1}^{N\times N_r} c_{I\sigma(I)}, \quad (15)$$

$$c_{IJ} = \frac{A^2(t_i,x_k)}{\tau^2}|t_i - t_j|^2 + \frac{A^2(t_i,x_k)}{\Delta x^2}|x_k - x_l|^2 + |d_1(t_i,x_k) - d_2(t_j,x_l)|^2, \quad (16)$$

with Δx an expected space-shift in the receiver dimension, and

$$I = i + (k-1)N, \;\; J = j + (l-1)N. \quad (17)$$

This new definition makes it possible to adapt the scaling to different parts of the shot gather to define the optimal assignment σ^* without inducing a renormalization of the adjoint source. This is important for the comparison of full seismic gathers, as the FWI misfit function writes in this case

$$\min_m f(m) = \sum_{s=1}^{N_s} \alpha_s \widehat{g}(d_{cal,s}, d_{obs,s}) \quad (18)$$

where $d_{cal,s}$ and $d_{obs,s}$ are 2D shot gathers. In such gathers, the amplitude dynamic from the zero offset trace to far offset trace is over several orders of magnitude, meaning that if A is chosen with respect to the maximum amplitude phase of the whole gather, the effective time/receiver shifts for smaller amplitude phases is rapidly small, leading to a distance measurement close from least-squares. Avoiding this effect requires at least trace-dependent scaling $A(x)$. The proposed GSOT formulation in this study makes it possible to use such trace-dependent scaling for computing the optimal assignment σ^* without inducing a trace-by-trace normalization which discards the AVO effect, and *which could not be compensated by the weight α_s*.

SECOND-ORDER DERIVATIVES AND PRECONDITIONING STRATEGIES

For the sake of concision, we consider here a single source/receiver couple and drop indices s and r. Consider $\sigma^*[m]$ is the solution of the assignment problem 2 between $d_{cal}[m]$ and d_{obs}. The GSOT misfit function can thus be written as

$$f(m) = \sum_{i=1}^{N} \frac{A^2}{\tau^2}|t_i - t_{\sigma^*(i)}|^2 + \left|d_{cal}[m](t_i) - d_{obs}\left(t_{\sigma^*[m](i)}\right)\right|^2. \quad (19)$$

One important result in Métivier et al. (2019) states that for a given m, the solution $\sigma^*[m]$ is almost everywhere unique (that is outside of point clouds configuration living in a space of codimension 1). From this uniqueness, the continuity of the distance function c_{ij} leads to the fact that $\sigma^*[m]$ is locally constant, and thus

$$\frac{\partial \sigma^*[m]}{\partial m} = 0, \;\; \text{almost everywhere.} \quad (20)$$

Therefore, we have

$$\nabla f(m) = J(m)\left(d_{cal}[m] - d_{obs}^{\sigma^*[m]}\right), \quad (21)$$

where $J(m)$ is the Jacobian operator such that

$$J(m) = \frac{\partial d_{cal}}{\partial m}. \quad (22)$$

Based on this result, the Hessian operator of the GSOT misfit function is simply given by

$$H(m) = J(m)^T J(m) + \frac{\partial J}{\partial m}\left(d_{cal}[m] - d_{obs}^{\sigma^*[m]}\right). \quad (23)$$

The latter equations shows that a Gauss-Newton approximation of the Hessian operator of the GSOT misfit function can be done. Based on this Gauss-Newton approximation, one can extract the diagonal terms of $J^T(m)J(m)$ and build a preconditioner $P(m)$ as the inverse of this diagonal, such that

$$P(m) = \text{diag}\left(J^T(m)J(m)\right)^{-1}. \quad (24)$$

Therefore, conventional preconditioners, such as pseudo-Hessian ones Choi and Shin (2008) designed for the acceleration of L^2 based FWI convergence can be used with equal efficiency in the frame of GSOT misfit function.

We consider a synthetic case study based on the P-wave velocity model from the Marmousi II benchmark model (Martin et al., 2006) (Fig. 2). We use a fixed spread acquisition with 96 sources from $x = 0.05$ km to $x = 12.59$ km each 132 m and 169 receivers from $x = 0.05$ km to $x = 16.85$ km each 100 m. The observed data is computed using a high-pass filtered Ricker function centered on 5 Hz such that there is no energy below 2.5 Hz. The density is kept constant. The simulations are run in the 2D acoustic approximation, using a 2nd-order in time and 4th- order in space finite-difference discretization on a uniform Cartesian grid with a 25 m discretization step.

The initial model is a 1D model linearly increasing with depth from the water layer to the bottom of the model (Fig. 2). We compare in Figures 3 the convergence of a L^2 based FWI and a GSOT based FWI with and without preconditioning both in terms of data and model errors. The final models obtained after 400 FWI iterations are presented in 4. As expected, due to cycle skipping, the L^2 based FWI converge towards a local minimum. The GSOT based inversion with no preconditioning converges towards the exact model but at a relatively slow pace. The pseudo-Hessian preconditioner significantly accelerates its convergence.

Finally, we also compare the l-BFGS strategy where the inverse Hessian is approximated from the l previous models and gradients (with $l = 20$ in this experiment) and a full BFGS strategy where the entire convergence history is used to build the inverse Hessian approximation. Interestingly, in these settings, the full BFGS strategy with pseudo-Hessian preconditioning provides the fastest convergence. This is an additional indication of the convexity of the GSOT misfit function. Were it be not convex, previous models and gradient should hardly contribute positively to the inverse Hessian approximation after a certain number of iterations.

CONCLUSION

GSOT is an interesting approach to mitigate cycle skipping in FWI. We present in this study its close link with DWT. The underlying misfit function can be seen as the solution of a regularized DWT problem over a space of specific time- shifts which can be related to a permutation of the time samples. We also analyze the scaling strategy on which relies GSOT, and propose an alternative scaling which yields a more flexible GSOT strategy, making it possible to introduce time and offset dependent expected time-shits without affecting the residuals amplitude. Finally, we demonstrate the simple form of the GSOT Hessian matrix, making it possible to use standard preconditioner such as pseudo-Hessian ones to accelerate the convergence. This is illustrated on the Marmousi model.

ACKNOWLEDGEMENTS

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

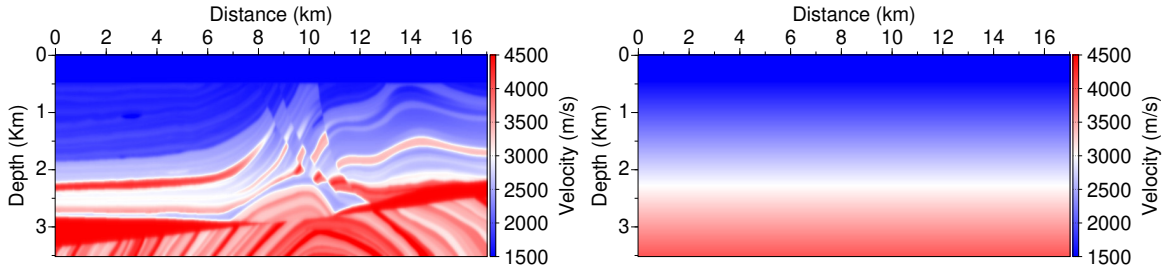

Figure 2: Exact (left) and initial (right) models.

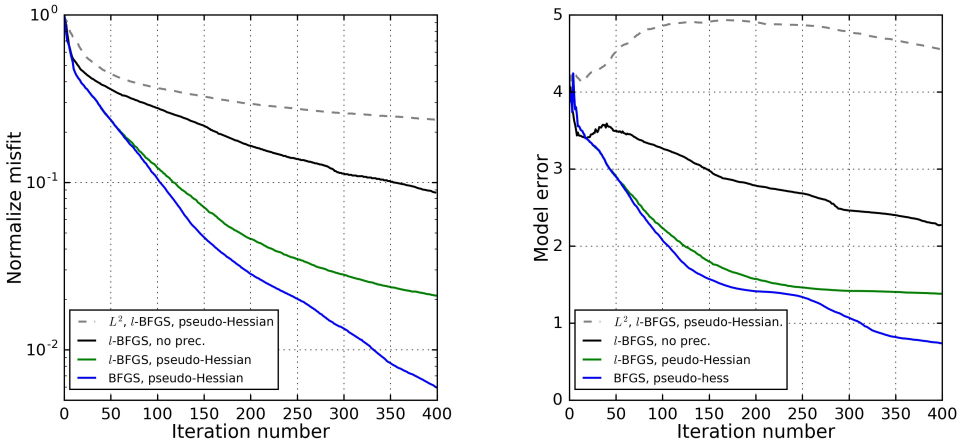

Figure 3: Convergence curves along FWI iterations. Normalized misfit functions (left), corresponding model error (right).

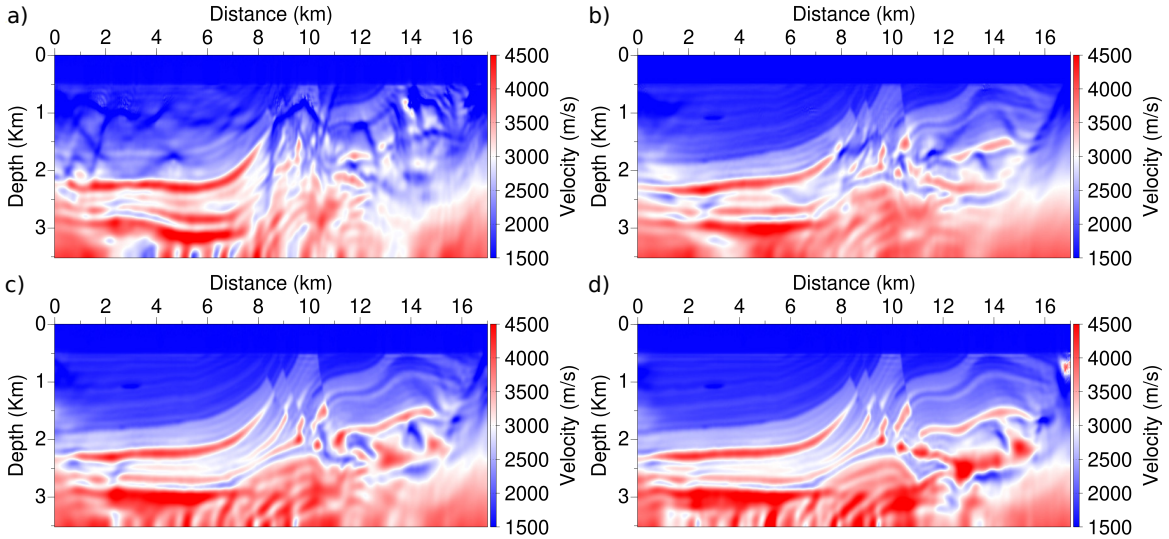

Figure 4: Final models after 400 iterations. L^2 inversion (a), l-BFGS GSOT inversion without preconditioning (b), l-BFGS GSOT inversion with pseudo-Hessian preconditioning (c), BFGS GSOT inversion with pseudo-Hessian preconditioning (d).

10.1190/segam2021-3583678.1

REFERENCES

Bertsekas, D. P., and D. Castanon, 1989, The auction algorithm for the transportation problem: Annals of Operations Research, **20**, 67–96, doi: https://doi.org/10.1007/BF02216923.

Billette, F., and G. Lambaré, 1998, Velocity macro-model estimation from seismic reflection data by stereotomography: Geophysical Journal International, **135**, 671–690, doi: https://doi.org/10.1046/j.1365-246X.1998.00632.x.

Choi, Y., and C. Shin, 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**, 2402–2415, doi: https://doi.org/10.1785/0120070179.

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

Górszczyk, A., R. Brossier, and L. Métivier, 2020, Graph-space optimal transport concept for time-domain FWI of OBS data: Nankai Trough velocity structure reconstructed from a 1D model: Journal of Geophysical Research, Solid Earth, **126**, e2020JB021504, doi: https://doi.org/10.1029/2020JB021504.

Jannane, M., W. Beydoun, E. Crase, D. Cao, Z. Koren, E. Landa, M. Mendes, A. Pica, M. Noble, G. Roeth, S. Singh, R. Snieder, A. Tarantola, and D. Trezeguet, 1989, Wavelengths of Earth structures that can be resolved from seismic reflection data: Geophysics, **54**, 906–910, doi: https://doi.org/10.1190/1.1442719.

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

Martin, G. S., R. Wiley, and K. J. Marfurt, 2006, Marmousi2: An elastic upgrade for Marmousi: The Leading Edge, **25**, 156–166, doi: https://doi.org/10.1190/1.2172306.

Métivier, L., R. Brossier, Q. Mérigot, and E. Oudet, 2019, A graph space optimal transport distance as a generalization of L p distances: Application to a seismic imaging inverse problem: Inverse Problems, **35**, 085001, doi: https://doi.org/10.1088/1361-6420/ab206f.

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

Pladys, A., R. Brossier, and L. Métivier, 2020, Graph space optimal transport based FWI: 3D OBC valhall case study: Presented at the SEG Technical Program, Expanded Abstracts, doi: https://doi.org/10.1190/segam2020-3420007.1.

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

Provenzano, G., R. Brossier, L. Metivier, and Y. Li, 2020, Joint FWI of diving and reflected waves using a graph space optimal transport distance: Synthetic tests on limited-offset surface seismic data: SEG International Exposition and Annual Meeting, 780–784, doi: https://doi.org/10.1190/segam2020-3426272.1.

Yang, Y., and B. Engquist, 2018, Analysis of optimal transport and related misfit functions in full-waveform inversion: Geophysics, **83**, no. 1, A7–A12, doi: https://doi.org/10.1190/geo2017-0264.1.