14th Biennial International Conference and Exposition
SPG 2023

Soft-dynamic time warping divergence in full-waveform inversion

Mahesh Kalita, Chris Purcell, and Lorenzo Casasanta*
Shearwater Geoservices
mkalita@shearwatergeo.com

Keywords

Full-waveform inversion, Cycle-skipping, Dynamic time warping

Abstract

Full-waveform inversion (FWI) seeks to find a subsurface model that minimizes the discrepancy between observed seismic data and modeled data. However, the commonly used L2-norm waveform difference misfit functional is susceptible to cycle-skipping, primarily due to the existence of local minima. To address this issue, we propose the utilization of the soft-dynamic time warping (SDTW) divergence distance as a more effective misfit metric in FWI. Our approach introduces a hyper-parameter in the formulation, which renders the functional differentiable. This enables the use of the adjoint state method to compute the gradient required for FWI. In contrast to the conventional SDTW, the divergence form of our proposed metric is always positive. The minimum value of the metric is achieved only when the modelled trace closely matches the observed trace. The application of the SDTW-divergence misfit in FWI framework to synthetic and field datasets demonstrates its robustness to cycle-skipping compared to the L2 norm.

Introduction

Full-waveform inversion (FWI) is a widely used technique in seismic exploration for obtaining accurate subsurface models. However, FWI often encounters difficulties in delivering geologically sound results, primarily due to the lack of low-frequency content and long offset distribution in the seismic data. These limitations, combined with a poor starting model, cause the nonlinear inversion process to converge to undesirable local minima of the misfit functional (Virieux & Operto, 2009). Consequently, numerous alternative cost functionals have been proposed over the past decades to replace the conventional L2 norm, aiming to maintain convexity even in the presence of significant model perturbations (Kalita & Alkhalifah, 2019).

In this study, we introduce a novel misfit functional for FWI based on the dynamic time warping (DTW) distance from modelled to observed seismic data. DTW is a well-known similarity measure that enables optimal matching between temporal sequences. Despite its popularity, DTW has rarely been utilized as a misfit functional in optimization problems such as FWI, primarily due to its non-differentiability (Cuturi & Blondel, 2017).

To address this challenge, Chen et al. (2022) recently proposed a differentiable version of DTW called soft-DTW. However, the soft-DTW-based misfit functional has two drawbacks. Firstly, it yields negative values, which is undesirable for FWI applications. Secondly, the minimum of the soft-DTW-based misfit functional does not necessarily occur when the predicted signal matches the observed signal. Motivated by the work of Blondel et al. (2021), we propose a solution to these issues by replacing soft-DTW with its divergence form. This modification ensures that the misfit functional, referred to as soft-DTW-divergence, overcomes the drawbacks of the original soft-DTW approach.

To demonstrate the versatility of the soft-DTW-divergence as an FWI misfit functional, we apply it to both synthetic and real data examples. For the synthetic data example, we rely on the infamous Marmousi model. Despite starting from a 1D model, the inversion results show that the soft-DTW-divergence significantly outperforms the conventional L2 norm misfit functional, producing a more accurate subsurface model estimation. In the real data example from an offshore Australia field, the soft-DTW-divergence again demonstrates its improved robustness to cycle-slipping and renders a geologically plausible subsurface model that generates synthetic waveforms that closely match the observed seismic data.

Theory

In general, DTW provides an optimum temporal alignment between two sets of time signals. To describe its methodology, let's consider a recorded seismic trace of n time samples, represented by $\boldsymbol{d} \in \mathbb{R}^n$, and its modelled counterpart by $\boldsymbol{p} \in \mathbb{R}^n$. We denote their elements by $\boldsymbol{d}_j, \boldsymbol{p}_i$ for $i, j \in [n]$, respectively. To calculate the DTW from $\boldsymbol{p}$ to $\boldsymbol{d}$, we need to find a set $\boldsymbol{\mathcal{A}}(n,n) \subset \{0,1\}^{n\times n}$ of all the possible monotonic binary alignment matrices $\boldsymbol{A}$, for which each element can be defined by:

$$[\boldsymbol{A}]_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{p_i} \text{ aligns with } \boldsymbol{d_j}, \\ 0, & \text{otherwise.} \end{cases}$$

While taking part in the temporal alignment procedure, every path $\boldsymbol{A}$ in $\boldsymbol{\mathcal{A}}$ incurs a cost. The path with the lowest cost is referred to as the DTW alignment path, and the cost is the DTW distance metric. Let $\boldsymbol{C}: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^{n\times n}$ be a function that maps $\boldsymbol{p}$ and $\boldsymbol{d}$ to a distance matrix. Without loss of generality, we consider the mapping function to be the Euclidian distance metric, which is defined by:

$$[\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d})]_{ij} = \frac{\boldsymbol{1}}{\boldsymbol{2}} || \boldsymbol{p}_i - \boldsymbol{d}_j ||_{\boldsymbol{2}}^{\boldsymbol{2}} .$$

The Frobenius inner product between $\boldsymbol{A}$ and $\boldsymbol{C}$, i.e., $< \boldsymbol{A}, \boldsymbol{C} > = \mathrm{tr}(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{A})$ is the sum of the costs along the alignment $\boldsymbol{A}$. Therefore, the DTW is the minimum cost among all the alignments:

$$DTW\ (\boldsymbol{C}) = \min_{\boldsymbol{A} \in \boldsymbol{\mathcal{A}}} < \boldsymbol{A}, \boldsymbol{C} > .$$

However, using the DTW distance as a misfit functional in an optimization problem such as FWI might produce a discontinuous gradient, especially when a small subsurface model perturbation causes a significant change in the optimal alignment matrix. This bottleneck mainly stems from the fact that the minimum operator, $\min(x_1, x_2, \dots, x_N)$, is not differentiable. To mitigate this issue, Cuturi & Blondel, 2017 propose to replace the hard definition of the minimum operator with its smooth version. They use the LogSumExp operator to define the smooth minimum approximation:

$$\min_\gamma (x_1, x_2, \dots, x_N) = -\gamma \log \sum_{i=1}^{N} e^{-\frac{x_i}{\gamma}},$$

where $\gamma > 0$ controls the trade-off between the approximation and smoothness. It is worth mentioning that $\min_\gamma$ converges to min as $\gamma \rightarrow 0$. The corresponding distance between $\boldsymbol{p}$ and $\boldsymbol{d}$ is named as soft-dynamic time warping, in notation, $\boldsymbol{SDTW_\gamma}(\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d}))$, i.e.,

$$SDTW\gamma\ (\boldsymbol{C}) = \min_{\boldsymbol{A} \in \boldsymbol{\mathcal{A}}}{}_\gamma < \boldsymbol{A}, \boldsymbol{C} > .$$

Chen, et al., 2022 first utilized $SDTW_\gamma$ in a non-linear FWI optimization framework. However, as pointed out in Blondel, et al., 2021, this misfit functional exhibits the following two properties:

1. There exists γ_0 such that $SDTW_\gamma(\boldsymbol{C}) \leq 0\ \forall\ \gamma \geq \gamma_0$.
2. The minimum of $SDTW_\gamma(\boldsymbol{C})$ is not achieved at $\boldsymbol{p} = \boldsymbol{d}$.

In this way, $SDTW_\gamma$ violates non-negativity and identity of indiscernible, two necessary conditions for a distance metric to be a statistical divergence. Its second property is especially problematic when employed as a misfit functional because it could drive FWI toward an optimal model where synthetic traces and field recordings are not required to match. To highlight this limitation, we consider a dummy seismic trace as the observed signal $\boldsymbol{d}$. Next, we mimic an ensemble of modelled signals $\boldsymbol{p}$, which are amplitude-scaled and positively or negatively time-shifted. We compute the misfit value ($SDTW_{\gamma=10}$) for different amplitude errors and time shift pairs, as in Figure 1a. The resulting functional is negative and it is not minimum when $\boldsymbol{p} = \boldsymbol{d}$ .To remove these limitations, Blondel, et al., 2021 proposes a derived metric including two correction terms, namely:

$$Div_\gamma^{\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d})} = SDTW_\gamma\big(\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d})\big) - SDTW_\gamma\big(\boldsymbol{C}(\boldsymbol{p},\boldsymbol{p})\big) - SDTW_\gamma\big(\boldsymbol{C}(\boldsymbol{d},\boldsymbol{d})\big).$$

As shown in Blondel, et al., 2021, the quantity $Div_\gamma^{\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d})} \geq 0$ for any $\boldsymbol{p}$ and $\boldsymbol{d}$, and $Div_\gamma^{\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d})} = 0$ only when $\boldsymbol{p} = \boldsymbol{d}$. Since Div_γ originates from $SDTW_\gamma$ and it follows the two necessary criteria of divergence

distance, it is called SDTW-divergence. Note that its limit tends to $DTW(\boldsymbol{C})$ as $\gamma \to 0$. When we substitute the Div_γ in the previous example, the minimum exists at $\boldsymbol{p} = \boldsymbol{d}$ as in Figure 1b.

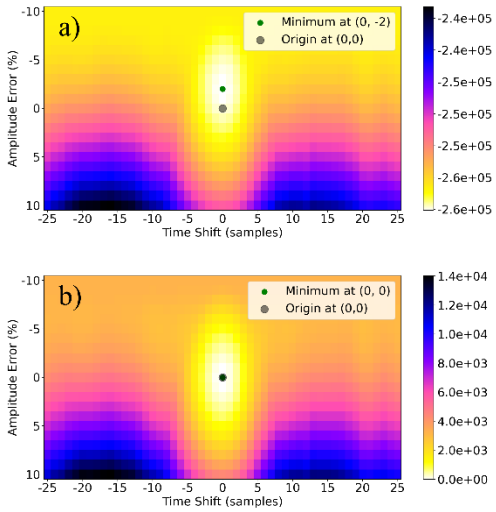

Figure 1: SDTW (a) and SDTW-Divergence (b) functionals evaluated in the presence of amplitude and time shift errors. Differently from SDTW, the SDTW-Divergence is a positive functional whose minimum is at $\boldsymbol{p} = \boldsymbol{d}$.

We suggest using the SDTW-divergence as an alternative misfit functional for FWI. Unlike the L2 norm, the SDTW-divergence maintains convexity for large model perturbations, reducing the issue of cycle-skipping into local minima. To demonstrate its convexity, we evaluate the misfit functional by sliding a dummy trace on itself. The L2 norm functional exhibits multiple local minima at different time shifts (Figure 2a), while the SDTW-divergence maintains convexity, particularly at low $\boldsymbol{\gamma}$ values (Figures 2b and 2c). However, low γ values introduce discontinuities in the misfit functional, making it unsuitable for gradient-based optimization methods like FWI. On the other hand, increasing γ make the functional differentiable, but at a cost of reducing convexity, leading to the appearance of local minima (Figure 2d). Therefore, the new functional is not completely immune to cycle-skipping, and its effectiveness depends on the tuning of hyperparameter γ.

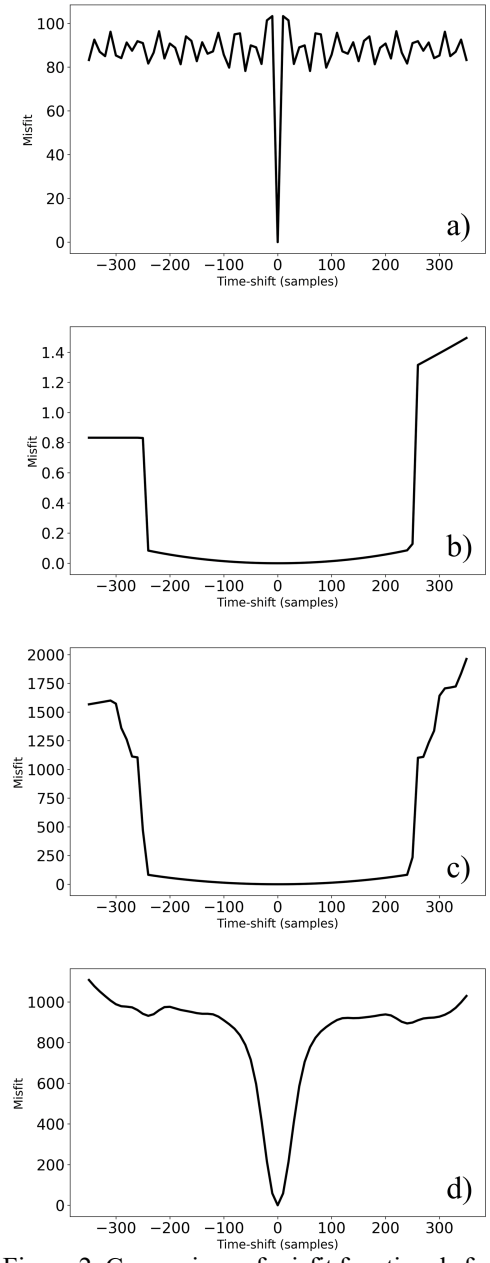

Figure 2: Comparison of misfit functionals for various time-shifts in a trace using (a) conventional L2 distance, (b) SDTW-Divergence with $\gamma = 0.001$, (c) $\gamma = 1$, and (d) $\gamma = 1000$. The use of lower γ values results in a non-differentiable functional, whereas higher values produce continuous derivatives but might introduce local minima.

To incorporate SDTW-divergence in a gradient-descent FWI inversion framework, we utilize the adjoint state method to calculate the gradient of $Div_{\gamma}^{\boldsymbol{C}(\boldsymbol{p}(\boldsymbol{m}),\boldsymbol{d})}$. The adjoint sources can be determined using the strategy outlined in Blondel, et al., 2021. In Figure 3, we observe the impact of the evaluation of adjoint sources extracted from the synthetic dataset example discussed in the next section. Figure 3a represents the conventional adjoint source i.e. $\boldsymbol{p}-\boldsymbol{d}$, while Figures 3b-3d display the adjoint sources of the proposed misfit functional for various γ values. As anticipated, small γ values result in unstable sources (Figure 3b), whereas larger γ values yield smoother adjoint sources (Figure 3d). This leads to smoother gradients and increases the chances of recovering large-scale features of the subsurface model. However, as mentioned earlier, fine-tuning the hyperparameter γ is crucial to achieving an optimal balance between model update smoothness and convexity of the functional.

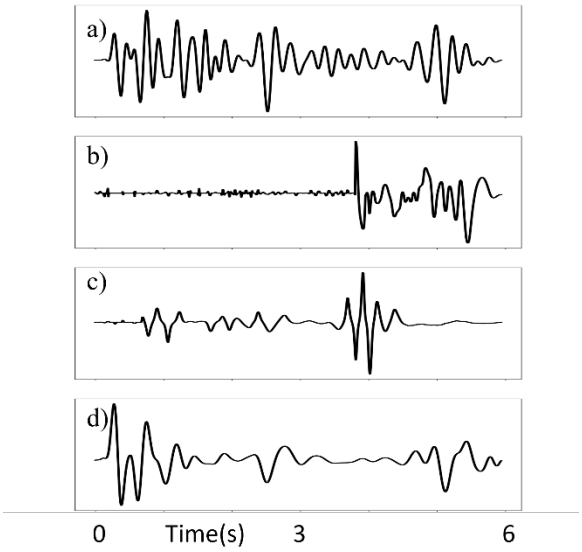

Figure 3: Comparison of adjoint sources using (a) conventional L2 norm, and SDTW-Divergence based FWI at (b) $\gamma = 0.001$, (c) $\gamma = 1$, (d) $\gamma = 1000$. A low value for γ makes the functional non-differentiable and consequently results in an unstable adjoint source .

Examples

Synthetic dataset: Figure 4a illustrates the target Marmousi model, which encompasses various geological features such as normal faults, tilted blocks towards the centre, intricate fine structures, and horizontal layering. To synthesize the observed data, a Ricker wavelet with a dominant frequency of 5 Hz is employed, considering the free surface boundary condition. The dataset is recorded at the sea surface using a fixed receiver spread spanning 8 km. We start

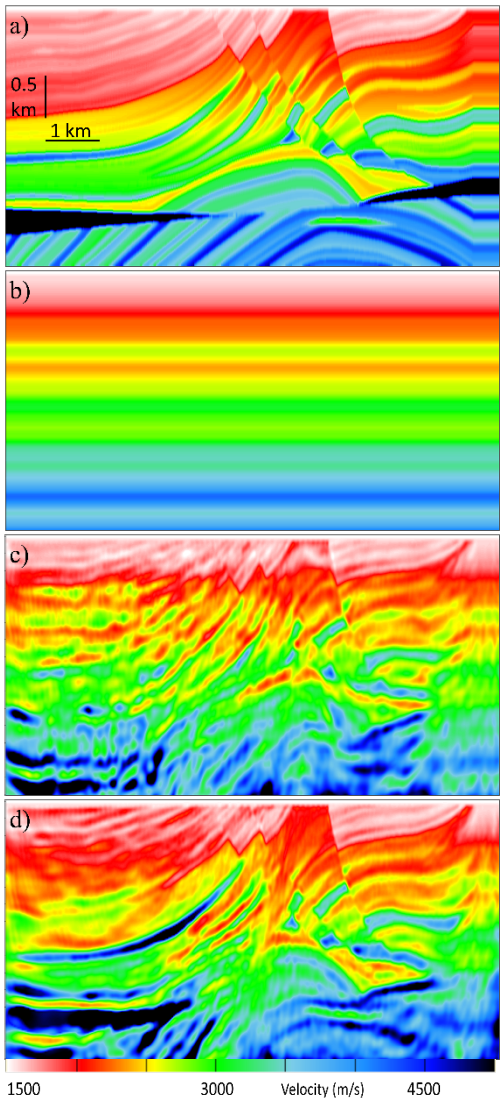

Figure 4: (a) Marmousi target model, (b) initial model, final inverted models using (c) classical L2 norm and (d) SDTW-divergence misfit functionals.

the inversion process with a 1-D velocity model, as depicted in Figure 4b. Following a hierarchical multiscale approach, the inversion process progresses from 3 Hz to 10 Hz. The final inverted models obtained using the conventional L2 and SDTW-divergence misfit functionals are displayed in Figures 4c and 4d, respectively. It is evident from the figures that the conventional method produces a result with cycle-skipping, while the SDTW-divergence approach yields a model that closely aligns with the ground truth model in Figure 4a.

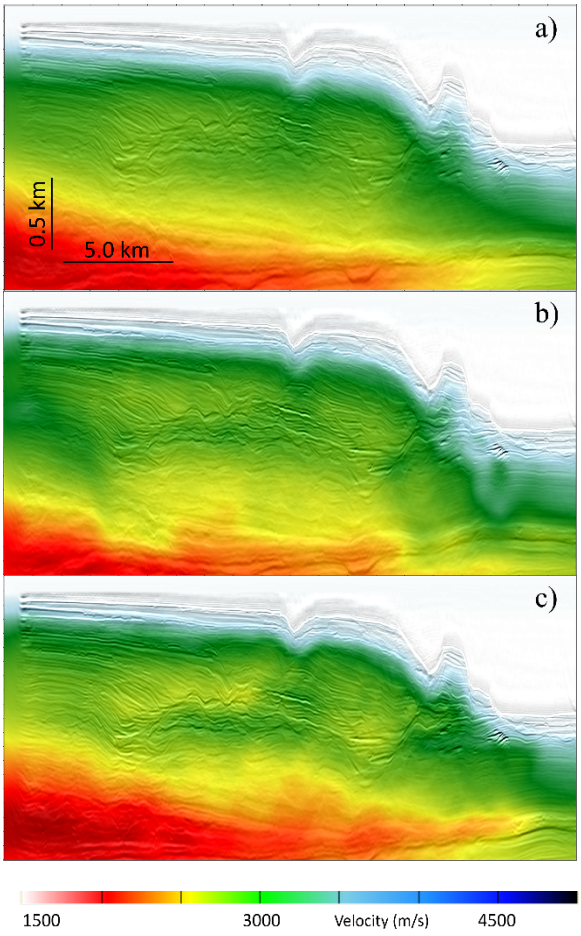

Figure 5: An inline slice from the (a) initial model, final inverted models using (b) classical L2 norm and (c) SDTW-divergence misfit functionals.

Field dataset: The 3d seismic data utilized in this study were obtained using a configuration of 8 streamers with a maximum offset of 4.6km. The streamers were approximately spaced at intervals of 100m, ensuring good subsurface illumination. The lowest usable frequency in the acquired data is approximately 4Hz. An analysis of diving waves shows they can penetrate up to 500–600 meters.

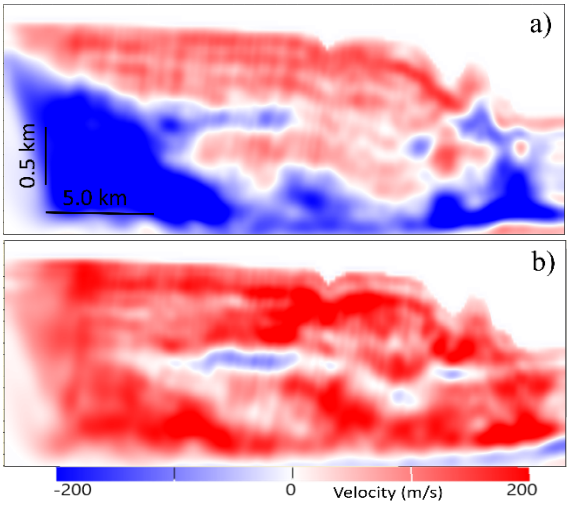

Figure 6: Cummumlative velocity updates using classical L2 norm (a) and SDTW-divergence (b) methods.

To create a challenging scenario for FWI, we deliberately designed the initial model to be slower than the available tomographic model. This choice is to accentuate the challenge of recovering positive values in the velocity field through the inversion process. Figure 5a displays the initial model superimposed with the migration image.

In this study, we conducted a total of 6 iterations of FWI focusing on a single frequency band up to 4Hz. Figures 5b and 5c display the final inverted models obtained using the conventional L2 norm and SDTW-divergence functional overlain on the imaged data with the equivalent model. It is evident that the latter SDTW method successfully recovers higher velocity values throughout the section compared to the conventional L2 norm. This improvement is further supported by Figure 6, which demonstrates that the cumulative FWI update using SDTW-divergence produces the anticipated overall model speed-up,

whereas the L2 norm updates clearly cycle-skips at depth and in some areas of the near-surface. Figure 7 displays a set of common image gathers calculated at an interval of 2 km to QC the updated models in the image domain. As anticipated, the events depicted in Figure 7a show an upward curvature, indicating that the initial velocity model is slower than the ground truth. In contrast, the final inverted model obtained using the SDTW-divergence method in Figure 5c successfully flattens the gathers (Figure 7c), while the inverted L2-norm model in Figure 5b fails to achieve the same level of flattening and focusing due to cycle skipping (Figure 7b).

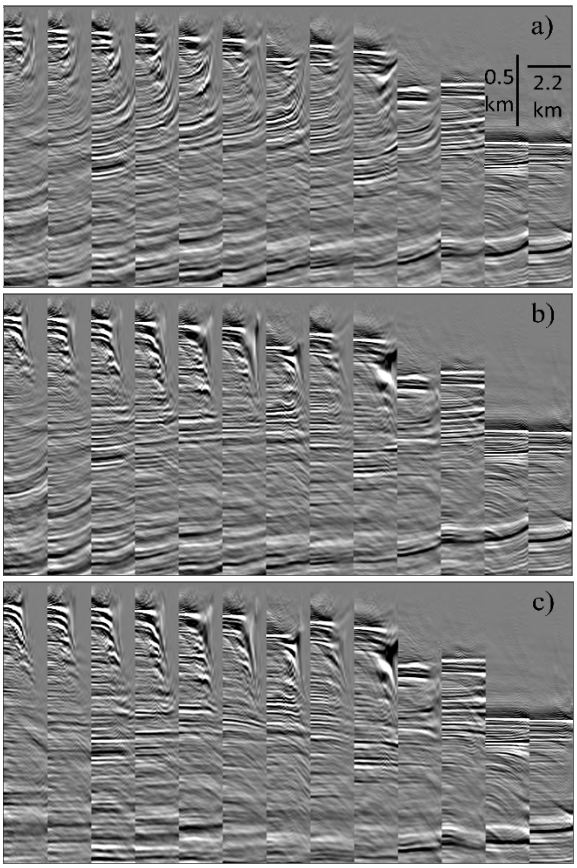

Figure 7: Common image gathers comparing initial (a) and final inverted models using the L2-norm (b) and SDTW-divergence functionalsproposed methods (c). Compared to the top two panels, the STDWT-divergence gathers show much-improved flattening and focusing. Note that these gathers have been extracted every 2km across the section.

Conclusions

We have introduced a novel objective functional for FWI that relies on the SDTW distance to compare predicted and observed waveforms. This newly devised formulation, known as SDTW-divergence, incorporates a hyper-parameter λ that ensures the functionals differentiability. Moreover, the SDTW-divergence functional is always positive and reaches its minimum when the predicted trace matches the observed one. However, caution should be exercised when selecting γ values. A larger γ value can lead to smoother updates but may increase the likelihood of cycle-skipping. On the other hand, a smaller γ value may result in sharper updates but could be less consistent as the functional approaches the non-differentiable regime. Finally, the FWI outcomes, utilizing both synthetic and field data, demonstrate that the SDTW-divergence outperforms the conventional L2 norm functional.

References

Blondel, M., Mensch, A. & Vert, J. P., 2021. Differentiable divergences between time series. *International conference on machine learning,* pp. 3853--3861.

Chen, F., Peter, D. & Ravasi, M., 2022. Cycle-skipping mitigation using misfit measurements based on differentiable dynamic time warping. *Geophysics,* 87(4), pp. R325--R335.

Cuturi, M. & Blondel, M., 2017. Soft-dtw: a differentiable loss function for time-series. *International conference on machine learning,* pp. 894--903.

Kalita, M. & Alkhalifah, T., 2019. Flux-corrected transport for full-waveform inversion. *Geophysical Journal International,,* 217(3), pp. 2147-2164.

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

Acknowledgements

We are grateful to Shearwater Geoservices for its permission to share this work. Thanks to Geoscience Australia for allowing us to use the field dataset. We thank Dr. Fuqiang Chen for his feedback on this work.