Differentiable dynamic-time warping divergences in full-waveform inversion

Mahesh Kalita* and Lorenzo Casasanta
Shearwater Geoservices

Summary

Full-waveform inversion (FWI) aims to obtain accurate subsurface models by minimizing the discrepancy between observed seismic data and modeled data. However, the commonly used L2-norm waveform difference misfit functional is prone to cycle-skipping due to local minima. To address this issue, we propose more effective misfit metrics for FWI by utilizing the soft-dynamic time warping (SDTW) divergence distance and its sharp variant. The proposed methods introduce a hyper-parameter to ensure differentiability of the functional, enabling the use of the adjoint state method for gradient computation in FWI. Unlike conventional SDTW, our divergence-based metrics always yield positive values, reaching their minimum when the modeled trace matches the observed trace. The efficacy of the proposed methods is demonstrated through their application on a field dataset, highlighting their robustness in mitigating cycle-skipping compared to the conventional L2 norm.

Introduction

Full-waveform inversion (FWI) is a widely used technique in seismic exploration for obtaining accurate subsurface models. However, it often encounters difficulties in delivering geologically sound results, primarily due to the lack of low-frequency content and inadequate offset distribution in the seismic data. These limitations, combined with a poor starting model, cause FWI to converge to undesirable local minima of the misfit functional (Virieux and Operto, 2009). Consequently, various alternative cost functionals have been proposed to preserve convexity in presence of large model perturbations. In this study, we introduce novel misfit functionals for FWI based on the dynamic time warping (DTW) distance from modeled to observed seismic traces.

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 and 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 may not be ideal for FWI applications. Secondly, a perfect match between the predicted and observed traces does not correspond to the minimum of the soft-DTW misfit functional, and this could cause bias during FWI iterations.

Motivated by the work of Blondel et al. (2021), we propose a solution 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. Furthermore, we remove the entropic term from the warping process to render another misfit functional, dubbed as sharp soft-DTW-divergence. To demonstrate the versatility of our proposed functionals, we consider a 3-D example from offshore Australia. Both functionals demonstrate improved robustness to cycle-skipping and render geologically plausible subsurface models that generate synthetic traces closely matching 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 \( d \in \mathbb{R}^n \), and its modeled counterpart by \( p \in \mathbb{R}^n \). We denote their elements by \( d_j, p_i \) for \( i, j \in [n] \), respectively. To calculate the DTW from \( p \) to \( d \), we need to find a set \( A(n, n) \subset \{0,1\}^{n \times n} \) of all the possible monotonic binary alignment matrices \( A \), for which each element can be defined by:

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

While taking part in the temporal alignment procedure, every path \( A \) in \( A \) 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 \( C : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n} \) be a function that maps \( p \) and \( d \) to a distance matrix. Without loss of generality, we consider the mapping function to be the Euclidean distance metric, which is defined by:

\[
[C(p, d)]_{ij} = \frac{1}{2} (p_i - d_j)^2.
\]

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

\[
DTW(C) = \min_{A \in A} \langle A, C \rangle.
\]

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, \( \min(x_1, x_2, ..., x_n) \), is not differentiable. To mitigate this issue, Cuturi and Blondel, 2017 propose to replace the hard
Differentiable DTW divergences in FWI

definition of the minimum operator with its smooth version. They use the LogSumExp operator to define the smooth minimum approximation:

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

where $\gamma > 0$ controls the trade-off between the approximation and smoothness. It is worth mentioning that $\min_{\gamma}$ converges to $\min$ as $\gamma \to 0$.

The corresponding distance between $p$ and $d$ is named as soft-dynamic time warping, in notation, $SDTW_{\gamma}(C(p,d))$, i.e.,

$$SDTW_{\gamma}(C) = \min_{A \in A} < A, C > .$$

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:

a) There exists $\gamma_0$ such that $SDTW_{\gamma}(C) \leq 0 \forall \gamma \geq \gamma_0$.

b) Minimum of $SDTW_{\gamma}(C)$ is not achieved at $p = d$.

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 $d$. Next, we mimic an ensemble of modeled signals $p$, 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 $p = d$. To remove these limitations, Blondel, et al., 2021 proposes a derived metric including two correction terms, namely:

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

As shown in Blondel, et al., 2021, the quantity $Div_{\gamma}^{C(p,d)} \geq 0$ for any $p$ and $d$, and $Div_{\gamma}^{C(p,d)} = 0$ only when $p = d$. Since $Div_{\gamma}$ 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 $DTW(C)$ as $\gamma \to 0$.

When we substitute the $Div_{\gamma}$ in the previous example, the minimum exists at $p = d$ 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 $p = d$.

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) $\gamma = 1000$. Note that higher $\gamma$ values can introduce local minima.
The variational form of $\text{SDTW}_\gamma$ involves decomposing the functional into two terms: a cost term and an entropy term ($H$). Mathematically, it can be expressed as

$$\text{SDTW}_\gamma (\mathbf{C}) = \langle E_\gamma(\mathbf{C}), \mathbf{C} \rangle > - \gamma H,$$

where $E_\gamma$ represents the expected alignment matrix that converges to the optimal alignment (DTW path) as $\gamma \rightarrow 0$. The cost term represents the directional derivative of soft-DTW in the direction of $\mathbf{C}$ and is referred to as “SHARP” due to its similarity to the concept proposed in optimal transport (Luise et al., 2018). While this version is inherently positive, it does not possess the property of indiscernibility. To address this, we propose the similar modification as mentioned above to make the metric achieve divergence:

$$S_\gamma(\mathbf{C}) = \text{SHARP}(\mathbf{C}(\mathbf{p}, \mathbf{d})) - \text{SHARP}(\mathbf{C}(\mathbf{p}, \mathbf{p})) - \text{SHARP}(\mathbf{C}(\mathbf{d}, \mathbf{d})).$$

We refer this modified term $S_\gamma(\mathbf{C})$ as sharp-soft-DTW-divergence.

![Figure 3](image)

Figure 3: Adjoint sources comparison for predicted and observed signals (a) using SDTW-divergence (b) and its sharp-version (c). The encircled region emphasizes subtle discrepancies between the two methods. Notably, a low $\gamma$ value renders the functional non-differentiable, leading to an unstable adjoint source.

We propose using the SDTW-divergence and its sharp variant as alternative misfit functionals for FWI. Unlike the L2 norm, the SDTW-divergence maintains convexity even for large model perturbations, mitigating cycle-skipping into local minima. The evaluation of the misfit functional by sliding a dummy trace on itself reveals the L2 norm’s multiple local minima (Figure 2a), while the SDTW-divergence maintains convexity, especially at low $\gamma$ values (Figure 2b). However, increasing $\gamma$ reduces convexity, leading to local minima (Figure 2c). Thus, while the new functional improves cycle-skipping, proper tuning of hyperparameter $\gamma$ remains crucial. The sharp variant exhibits similar curvature shapes to the divergent method but with different numerical values.

To incorporate SDTW-divergence and its sharp variant in a gradient-descent FWI inversion framework, we utilize the adjoint state method to calculate their associated gradients. The adjoint sources can be determined using the backward recursion strategy mentioned in Blondel, et al., 2021. Figure 3 shows the evaluation of adjoint sources of the proposed functionals for a pair of traces (Figure 3a) using different $\gamma$ values. Small $\gamma$ values result in unstable sources, whereas larger $\gamma$ values yield smoother adjoint sources. The two methods show a subtle difference, as highlighted in Figure 3c. Fine-tuning of $\gamma$ is crucial to achieving an optimal balance between model update smoothness and functional convexity.

Examples

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. In this study, we conducted a total of 6 iterations of FWI focusing on a single frequency band centered at 4Hz. Figure 4a displays the initial model superimposed with the migration image. Figures 4b-d display the final inverted models obtained using the conventional L2 norm and SDTW-divergence based functionals overlain on the imaged data with corresponding velocity models. It is evident that our proposed methods successfully recover higher velocity values throughout the section compared to the conventional L2 norm. The L2 norm updates clearly cycle-skips at depth and in some areas of the near-surface. This improvement is further supported by Figure 5, which 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 5a show an upward curvature, indicating that the initial velocity model is slower than the ground truth. In contrast, the final inverted models obtained using the SDTW-divergence based methods in Figures 4c-d successfully flatten the gatherings (Figures 5c-d), while the inverted L2-norm model in Figure 4b fails to achieve the same level of flattening and focusing due to cycle skipping (Figure 5b).
Conclusions

We propose novel objective functionals for FWI based on the SDTW-divergence distance, incorporating a hyperparameter $\gamma$ for differentiability. These functionals are always positive and minimize when predicted traces match observed ones. However, careful selection of $\gamma$ values is crucial: larger $\gamma$ values yield smoother updates but may increase cycle-skipping, while smaller $\gamma$ values lead to unstable adjoint sources and spurious model updates. FWI outcomes on a 3D field dataset demonstrate the superiority of our proposed functionals over the conventional L2 norm functional.

Acknowledgements

We would like to acknowledge Shearwater and our colleague, Chris Purcell, for their invaluable support. Additionally, we extend our appreciation to Dr. Fuqiang Chen for his valuable discussions and Geoscience Australia for providing the field dataset.