Differentiable Divergences Between Time Series

Mathieu Blondel
Google Research, Brain team

Arthur Mensch
École Normale Supérieure

Jean-Philippe Vert
Google Research, Brain team

Abstract

Computing the discrepancy between time series of variable sizes is notoriously challenging. While dynamic time warping (DTW) is popularly used for this purpose, it is not differentiable everywhere and is known to lead to bad local optima when used as a "loss". Soft-DTW addresses these issues, but it is not a positive definite divergence: due to the bias introduced by entropic regularization, it can be negative and it is not minimized when the time series are equal. We propose in this paper a new divergence, dubbed soft-DTW divergence, which aims to correct these issues. We study its properties; in particular, under conditions on the ground cost, we show that it is a valid divergence: it is non-negative and minimized if and only if the two time series are equal. We also propose a new "sharp" variant by further removing entropic bias. We showcase our divergences on time series averaging and demonstrate significant accuracy improvements compared to both DTW and soft-DTW on 84 time series classification datasets.

1 Introduction

Designing a meaningful discrepancy or "loss" between two sequences of variable lengths and integrating it in an end-to-end differentiable pipeline is challenging. For sequences on finite alphabets, differentiable local alignment kernels (Saigo et al., 2006) and edit distances (McCallum et al., 2012) have been proposed. For sequences on continuous domains, connectionist temporal classification (CTC) is popularly used in speech recognition (Graves et al., 2006). A related approach for time series motivated by geometry is dynamic time warping (DTW), which seeks a minimum-cost alignment between time series and can be computed by dynamic programming in quadratic time (Sakoe and Chiba, 1978). However, DTW is not differentiable everywhere, is sensitive to noise and is known to lead to bad local optima when used as a loss. Soft-DTW (Cuturi and Blondel, 2017) addresses these issues by replacing the minimum over alignments with a soft minimum, which has the effect of inducing a probability distribution over all alignments. Despite considering all alignments, it is shown that soft-DTW can still be computed by dynamic programming in the same complexity. Since then, soft-DTW has been successfully applied for audio to music score alignment (Mensch and Blondel, 2018), video segmentation (Chang et al., 2019), spatial-temporal sequences (Janati et al., 2020), and end-to-end differentiable text-to-speech synthesis (Donahue et al., 2020), to name but a few examples. Soft-DTW is included in popular R and Python packages for time series analysis (Sardá-Espinosa, 2017; Tavenard et al., 2020).

In this paper, we show that, despite recent successes, soft-DTW has some limitations which have been overlooked in the literature. First, it can be negative, which is a nuisance when used as a loss. Second, and more problematically, when used with a squared Euclidean cost, we show that it is never minimized when the two time series are equal. Put differently, given an input time series, the closest time series in the soft-DTW sense is never the input time series. This is due to the entropic bias introduced by replacing the minimum with a soft one. We propose in this paper a new divergence, dubbed soft-DTW divergence, which is based on soft-DTW but corrects for these issues. We study its properties; in particular, under condition on the ground cost, we show that it is a valid divergence: it is non-negative and it is minimized if and only if the two time series are equal. Our approach is related to Sinkhorn divergences (Ramdas et al., 2017; Genevay et al., 2018; Feydy et al., 2019), which use similar correction terms as we do for optimal transport distances, but our proof techniques are completely different. We also propose a new "sharp" variant by further

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

removing entropic bias. We showcase our divergences on time series averaging and demonstrate significant accuracy improvements compared to both DTW and soft-DTW on 84 time series classification datasets.

The rest of the paper is organized as follows. After reviewing some background in §2, we introduce the soft-DTW divergence and its "sharp" variant in §3. We study their properties and limit behavior. We study their empirical performance in §4 with experiments on time series averaging, interpolation and classification.

2 Background

2.1 Dynamic time warping

Let $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ be two d-dimensional time series of lengths m and n. We denote their elements by $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\boldsymbol{y}_j \in \mathbb{R}^d$, for $i \in [m]$ and $j \in [n]$. We say that $\boldsymbol{A} \in \{0,1\}^{m \times n}$ is an alignment matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$ when $[\boldsymbol{A}]_{i,j} = 1$ if $\boldsymbol{x}_i$ is aligned with $\boldsymbol{y}_j$ and 0 otherwise. We say that $\boldsymbol{A}$ is a monotonic alignment matrix if the ones in $\boldsymbol{A}$ form a path starting from the upper-left corner $(1,1)$ that connects the lower-right corner (m,n) using only $\downarrow$, $\rightarrow$, $\searrow$ moves. We denote the set of all such monotonic alignment matrices by $\mathcal{A}(m,n) \subset \{0,1\}^{m \times n}$. The cardinality $|\mathcal{A}(m,n)|$ grows exponentially in $\min(m,n)$ and is equal to the Delannoy number, $\text{Delannoy}(m-1, n-1)$, named after French amateur mathematician Henri Delannoy (Sulanke, 2003; Banderier and Schwer, 2005).

Let $C\colon \mathbb{R}^{m \times d} \times \mathbb{R}^{n \times d} \to \mathbb{R}^{m \times n}$ be a function which maps $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ to a distance or cost matrix $\boldsymbol{C} = C(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{m \times n}$. A popular choice is the squared Euclidean cost

$$[C(\boldsymbol{X}, \boldsymbol{Y})]_{i,j} = \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 \quad i \in [m], j \in [n]. \tag{1}$$

The Frobenius inner product $\langle \boldsymbol{A}, \boldsymbol{C} \rangle \coloneqq \text{Trace}(\boldsymbol{C}^\top \boldsymbol{A})$ between $\boldsymbol{C}$ and $\boldsymbol{A}$ is the sum of the costs along the alignment (Figure 1). Dynamic time warping (Sakoe and Chiba, 1978) can then be naturally formulated as the minimum cost among all possible alignments,

$$\textsc{dtw}(\boldsymbol{C}) \coloneqq \min_{\boldsymbol{A} \in \mathcal{A}(m,n)} \langle \boldsymbol{A}, \boldsymbol{C} \rangle. \tag{2}$$

The corresponding optimal alignment (not necessarily unique) is then

$$\boldsymbol{A}^\star(\boldsymbol{C}) \in \operatorname*{argmin}_{\boldsymbol{A} \in \mathcal{A}(m,n)} \langle \boldsymbol{A}, \boldsymbol{C} \rangle. \tag{3}$$

Despite the exponential number of alignments, (2) and (3) can be computed in $O(mn)$ time using dynamic programming and backtracking, respectively. The quantity $\textsc{dtw}(C(\boldsymbol{X}, \boldsymbol{Y}))$ is popularly used as a discrepancy measure between time series in numerous applications. In the rest of the paper, we will make the following assumptions about the ground cost C:

- A.1. $C(\boldsymbol{X}, \boldsymbol{Y}) \geq \boldsymbol{0}_{m \times n}$ (non-negativity),
- A.2. $[C(\boldsymbol{X}, \boldsymbol{X})]_{i,i} = 0$ for all $i \in [m]$,
- A.3. $C(\boldsymbol{X}, \boldsymbol{Y}) = C(\boldsymbol{Y}, \boldsymbol{X})^\top$ (symmetry).

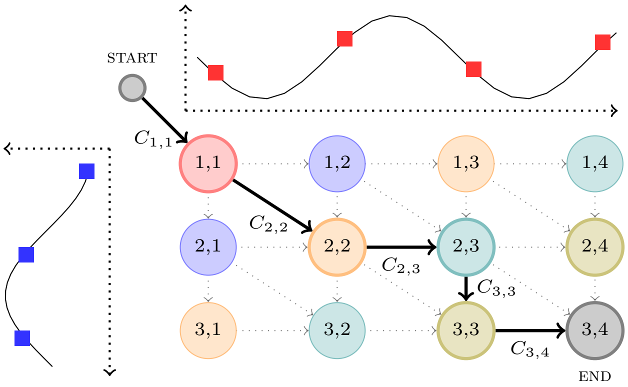

Figure 1: An alignment between two time series $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ corresponds to a path in a directed acyclic graph (DAG) and can be encoded as a binary matrix $\boldsymbol{A} \in \{0,1\}^{m \times n}$. The sum of the costs along the path is then $\langle \boldsymbol{A}, \boldsymbol{C} \rangle$. DTW seeks a minimum cost alignment, while soft-DTW seeks the soft minimum cost alignment. The latter induces a Gibbs distribution over all alignments.

The properties of $\textsc{dtw}$ under these assumptions are summarized in Table 1. Note that $\textsc{dtw}$ is minimized at $\boldsymbol{X} = \boldsymbol{Y}$ but this may not be the unique minimum.

2.2 Soft dynamic time warping

Definitions. In order to obtain a fully differentiable discrepancy measure between time series, Cuturi and Blondel (2017) proposed to replace the min operator in (2) by a smooth one,

$$\min_{x \in \mathcal{S}}{}_\gamma \; f(x) \coloneqq -\gamma \log \sum_{x \in \mathcal{S}} \exp(-f(x)/\gamma),$$

where $\gamma > 0$ is a parameter which controls the trade-off between approximation and smoothness. For convenience, we define the extension $\min_0 \coloneqq \min$. The resulting "soft" dynamic time warping formulation is

$$\begin{aligned} \textsc{sdtw}_\gamma(\boldsymbol{C}) &\coloneqq \min_{\boldsymbol{A} \in \mathcal{A}(m,n)}{}_\gamma \; \langle \boldsymbol{A}, \boldsymbol{C} \rangle \\ &= -\gamma \log \sum_{\boldsymbol{A} \in \mathcal{A}(m,n)} \exp(-\langle \boldsymbol{A}, \boldsymbol{C} \rangle / \gamma). \end{aligned} \tag{4}$$

Instead of only considering the minimum-cost alignment as in (2), (4) induces a Gibbs distribution over

Table 1: Properties of time-series losses under assumptions A.1-A.3 and differentiability of C. For the soft-DTW divergence, we prove non-negativity and "minimized at $\boldsymbol{X} = \boldsymbol{Y}$" using the cost (11) and one-dimensional absolute value (12) (cf. Proposition 3). For the soft-DTW and sharp divergences with the squared Euclidean cost (1), we only prove that $\boldsymbol{X} = \boldsymbol{Y}$ is a stationary point (cf. Proposition 4)

	Non-negativity	Minimized at $\boldsymbol{X} = \boldsymbol{Y}$	Symmetry	Differentiable everywhere
DTW	✓	✓	✓	×
Soft-DTW	×	×	✓	✓
Sharp soft-DTW	✓	×	✓	✓
Soft-DTW divergence	✓	✓	✓	✓
Sharp divergence	✓	✓	✓	✓
Mean-cost divergence	✓	✓	✓	✓

alignments. The probability of $\boldsymbol{A}$ given $\boldsymbol{C} \in \mathbb{R}^{m \times n}$ is

$$\mathbb{P}_\gamma(\boldsymbol{A}; \boldsymbol{C}) := \frac{\exp(-\langle \boldsymbol{A}, \boldsymbol{C} \rangle / \gamma)}{\sum_{\boldsymbol{A}' \in \mathcal{A}(m,n)} \langle -\langle \boldsymbol{A}', \boldsymbol{C} \rangle / \gamma)} \in (0, 1]. \quad (5)$$

We can see (4) as the negative log-partition of (5). For convenience, we also gather the probabilities of all possible alignments in a vector

$$\boldsymbol{p}_\gamma(\boldsymbol{C}) := (\mathbb{P}_\gamma(\boldsymbol{A}; \boldsymbol{C}))_{\boldsymbol{A} \in \mathcal{A}(m,n)} \in \triangle^{|A(m,n)|},$$

where $\triangle^k := \{\boldsymbol{p} \in \mathbb{R}^k \colon \boldsymbol{p} \geq \boldsymbol{0}_k, \boldsymbol{p}^\top \boldsymbol{1}_k = 1\}$ is the probability simplex. Let A be a random variable distributed according to (5). The expected alignment matrix under the Gibbs distribution induced by $\boldsymbol{C}$ is

$$\boldsymbol{E}_\gamma(\boldsymbol{C}) := \mathbb{E}_\gamma[A; \boldsymbol{C}] = \sum_{\boldsymbol{A} \in \mathcal{A}(m,n)} \mathbb{P}_\gamma(\boldsymbol{A}; \boldsymbol{C}) \boldsymbol{A} \in (0, 1]^{m \times n}. \quad (6)$$

Note that because the matrices in $\mathcal{A}(m,n)$ are binary ones, $[\boldsymbol{E}_\gamma(\boldsymbol{C})]_{i,j}$ is also equal to the marginal probability $\mathbb{P}_\gamma(A_{i,j} = 1; \boldsymbol{C})$, i.e., the probability that any of the paths goes through the cell (i, j).

Computation. Surprisingly, even though (4) contains a sum over all $\boldsymbol{A}$ in $\mathcal{A}(m,n)$, it can be computed in $O(mn)$ time by simply replacing the min operator with $\min_\gamma$ in the original dynamic programming recursion (Cuturi and Blondel, 2017). See also Algorithm 1 in Appendix A. The equivalence between (4) and this "locally smoothed" recursion was later formally proved using the associativity of the $\min_\gamma$ operator (Mensch and Blondel, 2018). The expected alignment can also be computed in $O(mn)$ time by backpropagation through the dynamic programming recursion (Cuturi and Blondel, 2017). See also Algorithm 2 in Appendix A.

Properties. The following proposition summarizes known properties of SDTW_γ (Cuturi and Blondel, 2017; Mensch and Blondel, 2018).

Proposition 1. *Properties of* SDTW_γ

The following properties hold for all $\boldsymbol{C} \in \mathbb{R}^{m \times n}$.

1. **Gradient:** $\text{SDTW}_\gamma(\boldsymbol{C})$ *is differentiable everywhere and its gradient is the expected alignment,*

$$\nabla_{\boldsymbol{C}} \text{SDTW}_\gamma(\boldsymbol{C}) = \boldsymbol{E}_\gamma(\boldsymbol{C}) \in (0, 1]^{m \times n}.$$

2. **Concavity:** $\text{SDTW}_\gamma(\boldsymbol{C})$ *is concave in* $\boldsymbol{C}$.
3. **Variational form:** *letting* $H(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} \rangle$,

$$\text{SDTW}_\gamma(\boldsymbol{C}) = \min_{\boldsymbol{p} \in \triangle^{|\mathcal{A}(m,n)|}} \langle \boldsymbol{p}, \boldsymbol{s}(\boldsymbol{C}) \rangle - \gamma H(\boldsymbol{p}) \quad (7)$$

where $\boldsymbol{s}(\boldsymbol{C}) := (\langle \boldsymbol{A}, \boldsymbol{C} \rangle)_{\boldsymbol{A} \in \mathcal{A}(mmn)} \in \mathbb{R}^{|\mathcal{A}(m,n)|}$.

4. **Scaling:** $\text{SDTW}_\gamma(\boldsymbol{C}) = \gamma \text{SDTW}_1(\boldsymbol{C}/\gamma)$, $\boldsymbol{E}_\gamma(\boldsymbol{C}) = \boldsymbol{E}_1(\boldsymbol{C}/\gamma)$ *and* $\boldsymbol{p}_\gamma(\boldsymbol{C}) = \boldsymbol{p}_1(\boldsymbol{C}/\gamma)$.
5. **Asymptotics:** $\text{DTW}(\boldsymbol{C}) \xleftarrow[0 \leftarrow \gamma]{} \text{SDTW}_\gamma(\boldsymbol{C})$ *and* $\boldsymbol{A}^\star(\boldsymbol{C}) \xleftarrow[0 \leftarrow \gamma]{} \boldsymbol{E}_\gamma(\boldsymbol{C})$.
6. **Lower and upper bounds:**

$$\text{DTW}(\boldsymbol{C}) - \gamma \log |\mathcal{A}(m,n)| \leq \text{SDTW}_\gamma(\boldsymbol{C}) \leq \text{DTW}(\boldsymbol{C}).$$

Note that $\text{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$ is generally neither convex nor concave in $\boldsymbol{X}$ and $\boldsymbol{Y}$, as is the case when C is the squared Euclidean cost (1). A notable exception is $C(\boldsymbol{X}, \boldsymbol{Y}) = -\boldsymbol{X}\boldsymbol{Y}^\top$, for which $\text{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$ is concave in $\boldsymbol{X}$ and $\boldsymbol{Y}$ (separately).

Use as a loss function. The differentiability of SDTW_γ makes it particularly suitable to use as a loss function between time series, of potentially variable lengths. An example of application is the computation of Fréchet means (1948) with respect to SDTW_γ. Specifically, given a set of k time series $\boldsymbol{Y}_1 \in \mathbb{R}^{n_1 \times d}$, $\ldots$, $\boldsymbol{Y}_k \in \mathbb{R}^{n_k \times d}$, we compute its average (barycenter) according to SDTW_γ by solving

$$\operatorname*{argmin}_{\boldsymbol{X} \in \mathbb{R}^{m \times d}} \sum_{i=1}^{k} w_i \ \mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}_i)), \tag{8}$$

where $\boldsymbol{w} = (w_1, \ldots, w_k) \in \mathbb{R}^k$ is a vector of pre-defined weights. When the time series $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$ have different lengths, a typical choice would be $w_i = 1/n_i$, to compensate for the fact that SDTW_γ increases with the length of the time series. Although it is non-convex, objective (8) can be solved approximately by gradient-based methods. Compared to DTW barycenter averaging (DBA) (Petitjean et al., 2011), it was shown that smoothing helps to avoid bad local optima. Using the chain rule and item 1 of Proposition 1, the gradient of $\mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$ w.r.t. $\boldsymbol{X}$ is

$$\nabla_{\boldsymbol{X}} \mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y})) = (J_{\boldsymbol{X}} C(\boldsymbol{X}, \boldsymbol{Y}))^\top \boldsymbol{E}_\gamma(\boldsymbol{C}(\boldsymbol{X}, \boldsymbol{Y})). \tag{9}$$

Here, we assume that C is differentiable and $J_{\boldsymbol{X}}$ denotes the Jacobian matrix of $\boldsymbol{C}(\boldsymbol{X}, \boldsymbol{Y})$ w.r.t. $\boldsymbol{X}$, a linear map from $\mathbb{R}^{m \times d}$ to $\mathbb{R}^{m \times n}$ (its transpose is a linear map from $\mathbb{R}^{m \times n}$ to $\mathbb{R}^{m \times d}$).

2.3 Global alignment kernel

Although it was introduced before soft dynamic time warping, the global alignment kernel (Cuturi et al., 2007) can be naturally expressed using SDTW_γ as

$$K_\gamma^C(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq \exp(-\mathrm{SDTW}_1(C(\boldsymbol{X}, \boldsymbol{Y})/\gamma)). \tag{10}$$

Using a constructive proof, it was shown that (10) is a positive definite (p.d.) kernel under certain cost functions and in particular with

$$[C(\boldsymbol{X}, \boldsymbol{Y})]_{i,j} = \delta(\boldsymbol{x}_i, \boldsymbol{y}_i) + \log(2 - \exp(-\delta(\boldsymbol{x}_i, \boldsymbol{y}_i)), \tag{11}$$

where $\delta(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. In the one-dimensional case ($d = 1$), we show in Appendix B.4 that

$$[C(\boldsymbol{X}, \boldsymbol{Y})]_{i,j} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_1, \tag{12}$$

also has the property that the kernel (10) is p.d. Using these costs, (10) can be used in any kernel method, such as support vector machines. The positive definiteness of (10) using the squared Euclidean cost (1) has to our knowledge not been proved or disproved yet.

3 New differentiable divergences

In this section, we begin by pointing out potential limitations of soft-DTW. We then introduce two new divergences, the soft-DTW divergence and its sharp variant, which aim to correct for these limitations. We study their properties and limit behavior.

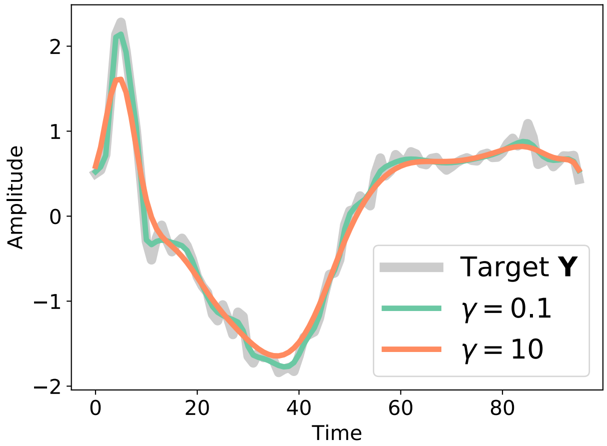

Figure 2: **Denoising effect of soft-DTW.** We show the result of $\mathrm{argmin}_{\boldsymbol{X}} \mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$, solved by L-BFGS with $\boldsymbol{X} = \boldsymbol{Y}$ as initialization, for two values of γ. As stated in Proposition 2, SDTW_γ with $\gamma > 0$ and squared Euclidean cost never achieves its minimum at $\boldsymbol{X} = \boldsymbol{Y}$. While this denoising can be useful, this means that SDTW_γ is not a valid divergence.

Limitations of soft-DTW. Despite recent empirical successes, soft-DTW has some inherent limitations that were not discussed in previous works. The following proposition clarifies these limitations.

Proposition 2. *Limitations of* SDTW_γ

The following holds.

1. *For all* $\boldsymbol{C} \in \mathbb{R}^{m \times n}$, $\gamma \mapsto \mathrm{SDTW}_\gamma(\boldsymbol{C})$ *is non-increasing, concave, and diverges to* $-\infty$ *when* $\gamma \to +\infty$. *In particular, there exists* $\gamma_0 \in [0, \infty)$ *such that* $\mathrm{SDTW}_\gamma(\boldsymbol{C}) \leq 0$ *for all* $\gamma \geq \gamma_0$.
2. *For all cost functions* C *satisfying A.2,* $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ *and* $\gamma \in [0, \infty)$, $\mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{X})) \leq 0$.
3. *For the squared Euclidean cost* (1) *and any* $\gamma \in (0, \infty)$, *the minimum of* $\mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$ *is not achieved at* $\boldsymbol{X} = \boldsymbol{Y}$.

A proof is given in Appendix B.3. Proposition 2 shows that that there exists values of γ or $\boldsymbol{C}$ for which $\mathrm{SDTW}_\gamma(\boldsymbol{C})$ is negative. Non-negativity is a useful property of divergences and the fact that SDTW_γ does not satisfy it can be a nuisance. More problematic is the fact that $\mathrm{SDTW}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$ is not minimized at $\boldsymbol{X} = \boldsymbol{Y}$. This is illustrated in Figure 2. While the denoising effect of soft-DTW can be useful, we would expect a proper differentiable divergence to be zero when the two time series are equal.

Soft-DTW divergences. To address these issues, we propose to use for all $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$

$$\begin{aligned} D_\gamma^C(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq\; & \textsc{sdtw}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y})) \\ & -\frac{1}{2}\textsc{sdtw}_\gamma(C(\boldsymbol{X}, \boldsymbol{X})) \\ & -\frac{1}{2}\textsc{sdtw}_\gamma(C(\boldsymbol{Y}, \boldsymbol{Y})). \end{aligned}$$

Since it is based on soft-DTW, we call it the soft-DTW divergence. Sinkhorn divergences (Ramdas et al., 2017; Genevay et al., 2018; Feydy et al., 2019), which are divergences between probability measures based on entropy-regularized optimal transport, use similar correction terms.

Sharp divergences. The variational form of $\textsc{sdtw}_\gamma$ (Proposition 1) implies that it can be decomposed as the sum of a cost term and an entropy term,

$$\textsc{sdtw}_\gamma(\boldsymbol{C}) = \langle \boldsymbol{E}_\gamma(\boldsymbol{C}), \boldsymbol{C} \rangle - \gamma H(\boldsymbol{p}_\gamma(\boldsymbol{C})). \tag{13}$$

On the other hand, we have

$$\textsc{dtw}(\boldsymbol{C}) = \langle \boldsymbol{A}^\star(\boldsymbol{C}), \boldsymbol{C} \rangle.$$

Since $\boldsymbol{E}_\gamma(\boldsymbol{C}) \to \boldsymbol{A}^\star(\boldsymbol{C})$ when $\gamma \to 0$, this suggests a new discrepancy measure,

$$\textsc{sharp}_\gamma(\boldsymbol{C}) \coloneqq \langle \boldsymbol{E}_\gamma(\boldsymbol{C}), \boldsymbol{C} \rangle. \tag{14}$$

It is the directional derivative of $\textsc{sdtw}_\gamma(\boldsymbol{C})$ in the direction of $\boldsymbol{C}$, since $\boldsymbol{E}_\gamma(\boldsymbol{C}) = \nabla_{\boldsymbol{C}} \textsc{sdtw}_\gamma(\boldsymbol{C})$. Inspired by Luise et al. (2018), who studied a similar idea in an optimal transport context,we call it sharp soft-DTW, since it removes the entropic regularization term $-\gamma H(\boldsymbol{p}_\gamma(\boldsymbol{C}))$ from (13). Its gradient is equal to

$$\nabla_{\boldsymbol{C}} \textsc{sharp}_\gamma(\boldsymbol{C}) = \boldsymbol{E}_\gamma(\boldsymbol{C}) + \frac{1}{\gamma} \nabla^2_{\boldsymbol{C}} \textsc{sdtw}_\gamma(\boldsymbol{C}) \boldsymbol{C} \in \mathbb{R}^{m \times n}, \tag{15}$$

where $\nabla^2_{\boldsymbol{C}} \textsc{sdtw}_\gamma(\boldsymbol{C}) \boldsymbol{C}$ is a Hessian-vector product (that can be computed efficiently, as we detail below). The gradient w.r.t. $\boldsymbol{X}$ is obtained by the chain rule, similarly to (9). Although $\textsc{sharp}_\gamma$ is trivially non-negative, it suffers from the same issue as $\textsc{sdtw}_\gamma$, namely, $\textsc{sharp}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y}))$ is not minimized at $\boldsymbol{X} = \boldsymbol{Y}$. We therefore propose to use instead

$$\begin{aligned} S_\gamma^C(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq\; & \textsc{sharp}_\gamma(C(\boldsymbol{X}, \boldsymbol{Y})) \\ & -\frac{1}{2}\textsc{sharp}_\gamma(C(\boldsymbol{X}, \boldsymbol{X})) \\ & -\frac{1}{2}\textsc{sharp}_\gamma(C(\boldsymbol{Y}, \boldsymbol{Y})). \end{aligned}$$

We call it the sharp soft-DTW divergence.

Validity. We remind the reader that in mathematics, a *divergence* D is a function that is non-negative ($D(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ for any $\boldsymbol{X}, \boldsymbol{Y}$) and that satisfies the identify of indiscernibles ($D(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X} = \boldsymbol{Y}$). By construction, we have $D_\gamma^C(\boldsymbol{X}, \boldsymbol{X}) = 0$ and $S_\gamma^C(\boldsymbol{X}, \boldsymbol{X}) = 0$ for all $\boldsymbol{X} \in \mathbb{R}^{m \times d}$. Moreover, the following result shows that D_γ^C is a valid divergence, under some assumptions on the cost C.

Proposition 3. *Valid divergence.*

Let $\gamma > 0$. If C is the cost defined in (11) *with $d \in \mathbb{N}$, or, if C is the absolute value* (12) *with $d = 1$, then $D_\gamma^C(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ for all $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$, and $D_\gamma^C(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X} = \boldsymbol{Y}$. Therefore, D_γ^C is a valid divergence.*

A proof is given in Appendix B.4. This implies that, for the costs (11) and (12), $D_\gamma^C(\boldsymbol{X}, \boldsymbol{Y})$ is uniquely minimized at $\boldsymbol{X} = \boldsymbol{Y}$. The proof relies on the fact that the global alignment kernel (10) is positive definite under these costs. Unfortunately, since the positive definiteness of (10) under the squared Euclidean cost (1) has not been proved or disproved, the same proof technique does not apply. Nevertheless, we can prove the following.

Proposition 4. *Stationary point under cost* (1)

If C is the squared Euclidean cost (1)*, then $\boldsymbol{X} = \boldsymbol{Y}$ is a stationary point of $D_\gamma^C(\boldsymbol{X}, \boldsymbol{Y})$ and $S_\gamma^C(\boldsymbol{X}, \boldsymbol{Y})$ w.r.t. $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ for all $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$.*

A proof is given in Appendix B.6. Based on Proposition 4 and ample numerical evidence (cf. Appendix B.5), we conjecture that $D_\gamma^C(\boldsymbol{X}, \boldsymbol{Y})$ and $S_\gamma^C(\boldsymbol{X}, \boldsymbol{Y})$ are also non-negative under the squared Euclidean cost.

Asymptotic behavior. We now study the behavior of our divergences in the zero and infinite temperature limits, i.e., when $\gamma \to 0$ and $\gamma \to \infty$. As we saw, $\boldsymbol{E}_\gamma(\boldsymbol{C})$ is the expected alignment matrix under the Gibbs distribution $\mathbb{P}_\gamma(\boldsymbol{A}; \boldsymbol{C})$. Let A be a random alignment matrix *uniformly* distributed over $\mathcal{A}(m, n)$, i.e., independent of the cost matrix $\boldsymbol{C}$. Replacing $\boldsymbol{E}_\gamma(\boldsymbol{C})$ with $\mathbb{E}[A]$ in (14), we obtain the mean cost, the average of the cost along all possible paths,

$$\begin{aligned} \textsc{mean\_cost}(\boldsymbol{C}) &\coloneqq \langle \mathbb{E}[A], \boldsymbol{C} \rangle \\ &= \frac{1}{|\mathcal{A}(m, n)|} \sum_{\boldsymbol{A} \in \mathcal{A}(m,n)} \langle \boldsymbol{A}, \boldsymbol{C} \rangle. \end{aligned} \tag{16}$$

We also define the mean-cost divergence,

$$\begin{aligned} M^C(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq\; & \textsc{mean\_cost}(C(\boldsymbol{X}, \boldsymbol{Y})) \\ & - \frac{1}{2}\textsc{mean\_cost}(C(\boldsymbol{X}, \boldsymbol{X})) \\ & - \frac{1}{2}\textsc{mean\_cost}(C(\boldsymbol{Y}, \boldsymbol{Y})). \end{aligned}$$

It bears some similarity with energy distances (Baringhaus and Franz, 2004; Székely et al., 2004), with the key difference that the probability distribution is over the alignments, not over the time series.

We now show that our proposed divergences are all intimately related through their asymptotic behavior, and that D^C_γ and S^C_γ share the same limits to the right when $m = n$ but not when $m \neq n$.

Proposition 5. *Limits w.r.t. γ*

For all $\boldsymbol{C} = C(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{m \times n}$, $m = n$:

$$\text{DTW}(\boldsymbol{C}) \xleftarrow[0 \leftarrow \gamma]{} D^C_\gamma(\boldsymbol{X}, \boldsymbol{Y}) \xrightarrow[\gamma \to \infty]{} M^C(\boldsymbol{X}, \boldsymbol{Y}).$$

For all $\boldsymbol{C} = C(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{m \times n}$, $m \neq n$:

$$\text{DTW}(\boldsymbol{C}) \xleftarrow[0 \leftarrow \gamma]{} D^C_\gamma(\boldsymbol{X}, \boldsymbol{Y}) \xrightarrow[\gamma \to \infty]{} \infty.$$

For all $\boldsymbol{C} = C(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{m \times n}$:

$$\text{DTW}(\boldsymbol{C}) \xleftarrow[0 \leftarrow \gamma]{} S^C_\gamma(\boldsymbol{X}, \boldsymbol{Y}) \xrightarrow[\gamma \to \infty]{} M^C(\boldsymbol{X}, \boldsymbol{Y}).$$

Note that the mean-cost divergence was obtained mostly as a side product of our limit case analysis. As we show in our experiments, it performs worse than the (sharp) soft-DTW divergence in practice. Therefore we do not recommend it in practice.

Computation. The value, gradient, directional derivative and Hessian product of $\text{SDTW}_\gamma(\boldsymbol{C})$ for $\boldsymbol{C} \in \mathbb{R}^{m \times n}$ can all be computed in $O(mn)$ time (Cuturi and Blondel, 2017; Mensch and Blondel, 2018). Therefore, both $D^C_\gamma(\boldsymbol{X}, \boldsymbol{Y})$ and $S^C_\gamma(\boldsymbol{X}, \boldsymbol{Y})$ take $O(\max\{m, n\}^2)$ time to compute. Sharp divergences take roughly twice more time to compute, as computing a Hessian-vector product requires one more pass through the dynamic programming recursion. The mean alignment and mean cost can also both be computed in $O(mn)$ time. We detail all algorithms in Appendix A.

Comparison with Sinkhorn divergences. Since our proposed divergences use similar correction terms as Sinkhorn divergences, we briefly review them and discuss their differences. Given two input probability measures $\boldsymbol{\alpha} \in \triangle^m$ and $\boldsymbol{\beta} \in \triangle^n$, entropy-regularized optimal transport is now commonly defined as

$$\text{OT}_\gamma(\boldsymbol{\alpha}, \boldsymbol{\beta}) \coloneqq \min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{C} \rangle + \gamma \text{KL}(\boldsymbol{T} || \boldsymbol{\alpha} \otimes \boldsymbol{\beta}), \quad (17)$$

where KL is the Kullback-Leibler divergence and $\mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the so-called transportation polytope (Peyré et al., 2019). To address the entropic bias of OT_γ, Sinkhorn divergences include correction terms, i.e., they are defined as $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto \text{OT}_\gamma(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \frac{1}{2}\text{OT}_\gamma(\boldsymbol{\alpha}, \boldsymbol{\alpha}) - \frac{1}{2}\text{OT}_\gamma(\boldsymbol{\beta}, \boldsymbol{\beta})$. There are however two important differences between OT_γ and $\text{SDTW}_\gamma(C(\cdot, \cdot))$. First, the former is convex in its inputs (separately) while the latter is not. This means that the proof technique for non-negativity of Sinkhorn divergences (Feydy et al., 2019) does not apply to the soft-DTW divergence. Indeed our proof technique for Proposition 3 is completely different than for Sinkhorn divergences. Second, the entropic regularization in SDTW_γ is on the probability distribution (Proposition 1), not on the soft alignment, as is the case for the transportation map $\boldsymbol{T}$ in (17). Contrary to Sinkhorn divergences, the soft-DTW and sharp divergences are non-convex in their inputs. For time-series averaging, an initialization scheme that works well in practice is to use the SDTW_γ solution as initialization, itself initialized from the Euclidean mean.

4 Experimental results

Throughout this section, we use the UCR (University of California, Riverside) time series classification archive (Chen et al., 2015). We use a subset containing 84 datasets encompassing a wide variety of fields (astronomy, geology, medical imaging) and lengths. Datasets include class information (up to 60 classes) for each time series and are split into train and test sets. Due to the large number of datasets in the UCR archive, we choose to report only a summary of our results in the main manuscript. Detailed results are included in the appendix for interested readers. In all experiments, we use the squared Euclidean cost (1). Our Python source code is available on github.

4.1 Time series averaging

Experimental setup. To investigate the effect of our divergences on time series averaging, we replace SDTW_γ in objective (8) with our divergences. For this task, we focus on a visual comparison and refrain from reporting quantitative results, since the choice of evaluation metric necessarily favors one divergence over others. For each dataset, we pick 10 time series $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{10}$ randomly. Since the time series all have the same length, we use uniform weights $w_1 = \cdots = w_k = 1$. To approximately minimize the objective function, we use 200 iterations of L-BFGS (Liu and Nocedal, 1989). Because the objective is non-convex in $\boldsymbol{X}$, initialization is important. For DTW, SDTW_γ, SHARP_γ and MEAN_COST, we use the Euclidean mean as initialization and set $\gamma = 1$. For D^C_γ, S^C_γ and M^C, we use as initialization the solution of their "biased couterpart", i.e., SDTW_γ, SHARP_γ,

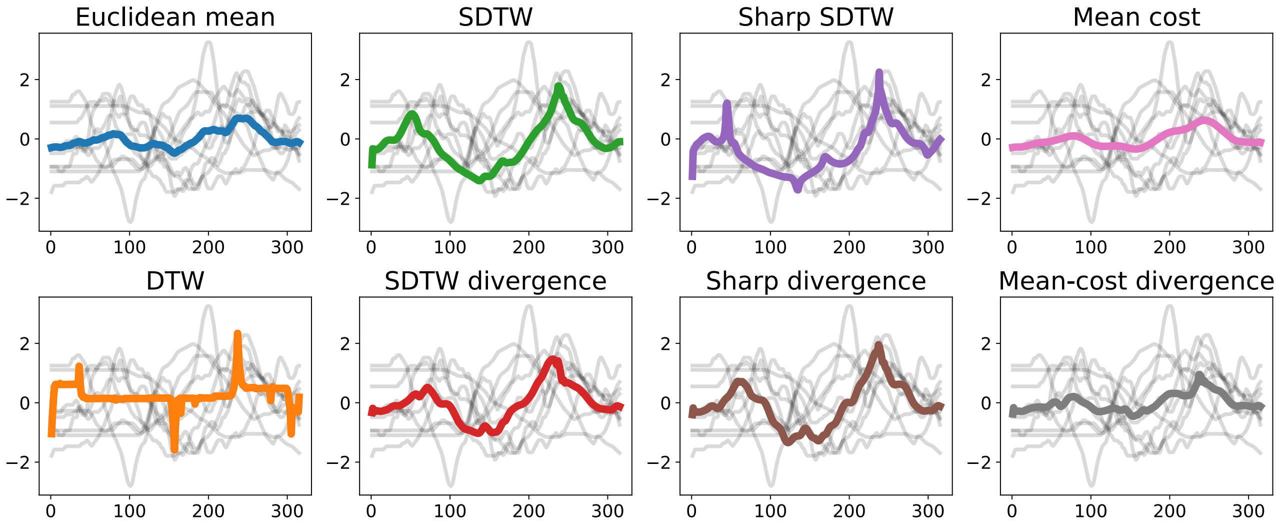

Figure 3: Average of 10 time series $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{10}$, on the **uWaveGestureLibrary_Y** dataset.

MEAN_COST, respectively, and we set $\gamma = 10$.

Results. We show the time series averages obtained on the *uWaveGestureLibrary_Y* dataset in Figure 3. With DTW, the obtained average does not match well the time series, confirming the conclusion of Cuturi and Blondel (2017). This is because the objective is both highly non-convex and non-smooth, rendering optimization difficult, despite the use of Euclidean mean as initialization. On the other hand, the averages obtained by other divergences appear to match the time series much better, thanks to the smoothness of their objective function. We observe that D_γ^C (soft-DTW divergence), S_γ^C (sharp divergence) and M^C (mean-cost divergence) produce different results from their biased counterpart, SDTW$_\gamma$ (soft-DTW), SHARP$_\gamma$ (sharp soft-DTW) and MEAN_COST (mean cost), respectively. This is to be expected, since the variable $\boldsymbol{X}$ with respect to which we minimize is involved in the correcting term using $C(\boldsymbol{X}, \boldsymbol{X})$. The averages obtained with SHARP$_\gamma$ and S_γ^C tend to include sharper peaks, a trend confirmed on other datasets as well. More average examples are included in the appendix.

4.2 Time series interpolation

Experimental setup. As a simple variation of time series averaging, we now consider time series interpolation. We pick two times series $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ and set the weights in objective (8) to $w_1 = \pi$ and $w_2 = 1 - \pi$, for $\pi \in \{0.25, 0.5, 0.75\}$, i.e., we seek an interpolation of the two time series. We again minimize the objective approximately using L-BFGS, with the same initialization scheme and the same γ as before.

Results. Results on the *ArrowHead* dataset are shown in Figure 4. We observe similar trends as for time series averaging. The interpolations obtained by DTW include artifacts that do not represent well the data. Our divergences obtain slightly more visually pleasing results than their biased counterparts. More examples are included in the appendix. The interpolation obtained by the sharp soft-DTW includes a peak (light green) which is slightly off, but this is not the case of the sharp divergence.

4.3 Time series classification

Experimental setup. To quantitatively compare our proposed divergences, we now consider time series classification tasks. To better isolate the effect of the divergence itself, we choose two simple classifiers: nearest neighbor and nearest centroid. To predict the class of a time series, the well-known nearest neighbor classifier assigns the class of the nearest time series in the training set, according to the chosen divergence. Note that this does not require differentiability of the divergence. The lesser known nearest centroid classifier (Hastie et al., 2001) first computes the centroid (average) of each class in the training set. We compute the centroid by minimizing (8) for each class, according to the chosen divergence. To predict the class of a time series, we then assign the class of the nearest centroid, according to the same divergence. Although very simple, this method is known to be competitive with the nearest neighbor classifier, while requiring much lower computational cost at prediction time (Petitjean et al., 2014).

For all datasets in the UCR archive, we use the predefined test set. For divergences including a γ param-

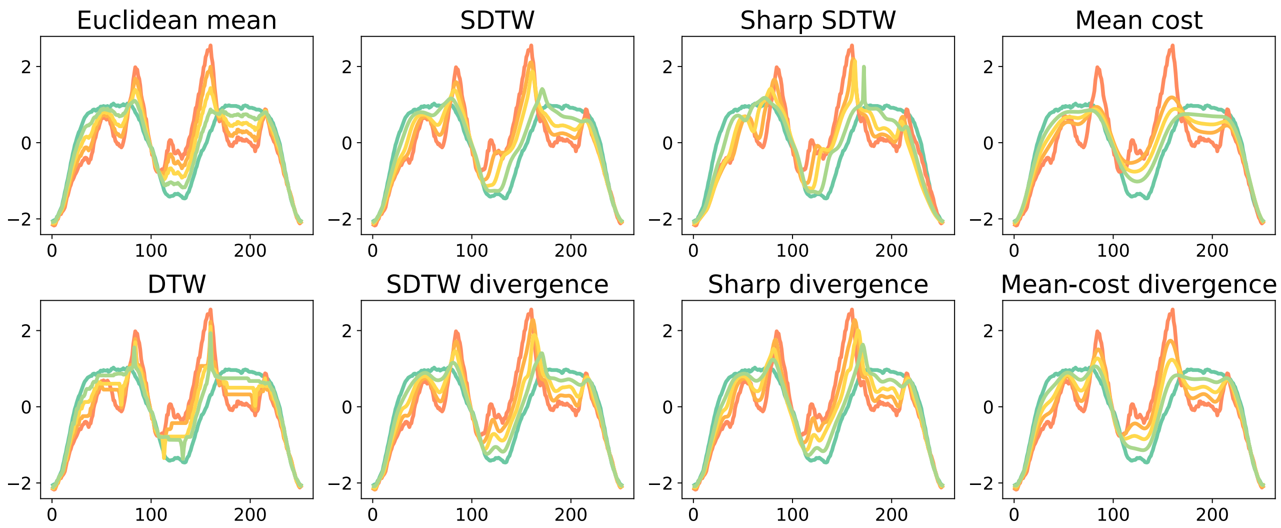

Figure 4: Interpolation between two time series $\boldsymbol{Y}_1$ (red) and $\boldsymbol{Y}_2$ (dark green), from the **ArrowHead** dataset.

eter, we select γ by cross-validation. More precisely, we train on 2/3 of the training set and evaluate the goodness of a γ value on the held-out 1/3. We repeat this procedure 5 times, each with a different random split, in order to get a better estimate of the goodness of γ. We do so for $\gamma \in \{10^{-4}, 10^{-3}, \dots, 10^{4}\}$ and select the best one. Finally, we retrain on the entire training set using that γ value.

Results. Due to the large number of datasets in the UCR archive, we only show a summary of the results in Table 2 and Table 3. Detailed results are in Appendix C. We observe consistent trends for both the nearest neighbor and the nearest centroid classifiers. The mean-cost divergence appears to perform poorly, even worse than the squared Euclidean distance and DTW. This shows that considering all possible alignments uniformly does not lead to a good divergence measure. On the other hand, our proposed divergences, the soft-DTW divergence and the sharp divergence, outperform on the majority of the datasets the Euclidean distance, DTW, soft-DTW, and sharp soft-DTW. Furthermore, each proposed divergence (i.e., with correction term) clearly outperforms its biased counterpart (i.e., without correction term). This shows that proper divergences, which are minimized when the two time series are equal, indeed translate to higher classification accuracy in practice. Overall, the soft-DTW divergence works better than the sharp divergence.

5 Conclusion

Due to entropic bias, soft-DTW can be negative and is not minimized when the two time series are equal. To address these issues, we proposed the soft-DTW divergence and its sharp variant. We proved that the former is a valid divergence under the cost (11) for $d \in \mathbb{N}$ and under the absolute cost (12) for $d = 1$. We conjecture that this is also true under the squared Euclidean cost (1), but leave a proof to future work. By studying the limit behavior of our divergences when the regularization parameter γ goes to infinity, we also obtained a new mean-cost divergence, which is of independent interest. Experiments on 84 time series classification datasets established that the soft-DTW divergence performs the best among all discrepancies and divergences considered.

References

Cyril Banderier and Sylviane Schwer. Why Delannoy numbers? *Journal of statistical planning and inference*, 135(1):40–54, 2005.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proc. of CVPR*, pages 3546–3555, 2019.

Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proc. of ICML*, 2017.

Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes,

Table 2: **Nearest neighbor results.** Each number indicates the percentage of datasets in the UCR archive for which using A in the nearest neighbor classifier is within 99% or better than using B .

A (↓) vs. B (→)	Euc.	DTW	SDTW	SDTW div	Sharp	Sharp div	Mean cost	Mean-cost div
Euclidean	-	41.67	34.62	22.37	29.49	27.63	95.29	71.43
DTW	71.43	-	42.31	39.47	50.00	39.47	89.29	79.76
SDTW	75.64	82.05	-	52.63	73.08	55.26	97.44	80.77
SDTW div	93.42	93.42	86.84	-	84.21	82.67	97.37	96.05
Sharp	83.33	84.62	76.92	53.95	-	52.63	98.72	87.18
Sharp div	94.74	86.84	77.63	66.67	81.58	-	98.68	96.05
Mean cost	9.41	13.10	8.97	5.26	5.13	6.58	-	44.05
Mean-cost div	42.86	32.14	25.64	19.74	21.79	18.42	98.81	-

Table 3: **Nearest centroid results.** Each number indicates the percentage of datasets in the UCR archive for which using A in the nearest neighbor classifier is within 99% or better than using B .

A (↓) vs. B (→)	Euc.	DTW	SDTW	SDTW div	Sharp	Sharp div	Mean cost	Mean-cost div
Euclidean	-	44.71	27.06	28.57	30.95	32.50	77.65	78.82
DTW	63.53	-	36.47	36.90	41.67	37.50	83.53	80.00
SDTW	82.35	85.88	-	55.95	77.38	62.50	94.12	94.12
SDTW div	82.14	83.33	82.14	-	78.57	70.00	91.67	94.05
Sharp	79.76	78.57	54.76	48.81	-	55.00	91.67	91.67
Sharp div	82.50	82.50	70.00	63.75	78.75	-	92.50	93.75
Mean cost	37.65	22.35	11.76	11.90	15.48	11.25	-	77.65
Mean-cost div	41.18	23.53	14.12	14.29	17.86	15.00	90.59	-

and Tomoko Matsui. A kernel for time series based on global alignments. In *Proc. of ICASSP*, volume 2, pages II–413. IEEE, 2007.

Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575*, 2020.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310. Presses universitaires de France, 1948.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of ICML*, pages 369–376, 2006.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer New York Inc., 2001.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In *Proc. of AISTATS*, pages 1695–1704. PMLR, 2020.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Proc. of NeurIPS*, pages 5859–5870, 2018.

Andrew McCallum, Kedar Bellare, and Fernando Pereira. A conditional random field for discriminatively-trained finite-state string edit distance. *arXiv preprint arXiv:1207.1406*, 2012.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proc. of ICML*, 2018.

François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011.

François Petitjean, Germain Forestier, Geoffrey I Webb, Ann E Nicholson, Yanping Chen, and Eamonn Keogh. Dynamic time warping averaging of time series allows faster and more accurate classification. In *ICDM*, pages 470–479. IEEE, 2014.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11 (5-6):355–607, 2019.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19 (2):47, 2017.

Hiroto Saigo, Jean-Philippe Vert, and Tatsuya Akutsu. Optimizing amino acid substitution matrices with a local alignment kernel. *BMC bioinformatics*, 7(1):246, 2006.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Alexis Sardá-Espinosa. Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, 12:41, 2017.

Robert A Sulanke. Objects counted by the central delannoy numbers. *J. Integer Seq*, 6(1), 2003.

Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5 (16.10):1249–1272, 2004.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *JMLR*, 21(118):1–6, 2020.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.