Aligning Time Series on Incomparable Spaces

Samuel Cohen
Centre for Artificial Intelligence
University College London

Giulia Luise
Department of Computing
Imperial College London

Alexander Terenin
Department of Mathematics
Imperial College London

Brandon Amos
Facebook AI Research

Marc Peter Deisenroth
Centre for Artificial Intelligence
University College London

Abstract

Dynamic time warping (DTW) is a useful method for aligning, comparing and combining time series, but it requires them to live in comparable spaces. In this work, we consider a setting in which time series live on different spaces without a sensible ground metric, causing DTW to become ill-defined. To alleviate this, we propose Gromov dynamic time warping (GDTW), a distance between time series on potentially incomparable spaces that avoids the comparability requirement by instead considering intra-relational geometry. We demonstrate its effectiveness at aligning, combining and comparing time series living on incomparable spaces. We further propose a smoothed version of GDTW as a differentiable loss and assess its properties in a variety of settings, including barycentric averaging, generative modeling and imitation learning.

1 Introduction

Data is often gathered sequentially in the form of a time series, which consists of a sequence of data points observed at successive time points. Elements of such sequences are correlated through time, and comparing time series requires one to take the direction of time into account. To define a meaningful similarity measure between time series, Sakoe and Chiba (1978) proposed *dynamic time warping* (DTW), a distance over the space of time series. DTW consists of a minimal-cost alignment problem and is solved efficiently via dynamic programming.

Dynamic time warping enables one to tackle a large range of temporal problems, including aligning, comparing, and averaging time series. In particular, DTW can be employed as a loss function within larger learning frameworks: in this setting, Cuturi and Blondel (2017) propose *soft DTW*, which consists of a smoothed DTW objective possessing a differentiable gradient which can result in better behavior when employing gradient-based methods (Cuturi, 2011).

DTW and its variants require a sensible cost function to be defined between samples from the two time series. The specification of such cost functions is often hard, and limits the applicability of DTW. For example, in cases where the time series are invariant under symmetries, such as sequences of word embeddings which are only identified up to a rotation of latent space, one needs to solve a spatial alignment problem to compare the two sequences sensibly.

Zhou and Torre (2009) propose an extension of DTW that addresses this issue by jointly optimizing spatial and temporal projections that align the time series. Vayer et al. (2020) introduce a similar extension of DTW that consists in making the cost invariant with respect to specific sets of invariances, such as for example rotations. In these approaches, one still requires the definition of a cost function between samples from the two time series, along with a potentially large predefined set of transformations to optimize over. On the other hand, in multi-modal settings, one considers time series that live on incomparable spaces: for ex-

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).

Code available at: HTTPS://GITHUB.COM/SAMCOHEN16/ALIGNING-TIME-SERIES.

ample, the configuration space of a robotic arm and its representation as pixels of a video frame. In such cases, defining a sensible distance between different representations and a sensible space of symmetries is impractical, as it would require detailed understanding of the objects we wish to study.

In this work, we propose to tackle the incomparability and invariance problems simultaneously by relaxing our notion of equality in a manner inspired by recent ideas from the optimal transport literature. Using connections between DTW and the Wasserstein distance (Kantorovich, 1958), we propose *Gromov dynamic time warping* (GDTW), which compares two time series by contrasting their intra-relational geometries, analogously to the Gromov–Wasserstein distance of isometry classes of metric-measure spaces (Mémoli, 2011). This allows one to compare two time series without requiring a similarity notion between their samples. The resulting procedure automatically incorporates invariances into the distance, without requiring said invariances or symmetry-specific constraints to be manually specified.

Contributions. (1) We introduce a new distance between time series that is well-defined on incomparable spaces with naturally built-in invariance to isometries, and (2) a smoothed extension with better-behaved gradients. (3) We propose an efficient Frank–Wolfe-inspired algorithm for computing it, and (4) we apply Gromov DTW as a loss function in a wide range of settings, including barycentric averaging, generative modeling and imitation learning.

Notation. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a compact metric space, and let a *time series* $\boldsymbol{x}$ of length $T \in \mathbb{N}$ be an element of $\mathcal{X}^T$. Let $\mathcal{A}(m,n) \subseteq \{0,1\}^{m\times n}$ be the set of *alignment matrices*, which are binary matrices containing a path of ones from the top-left to the bottom-right corner, allowing only bottom, right or diagonal bottom-right moves. Given a matrix $\mathbf{A} \in \mathcal{A}(m,n)$ and a 4-dimensional array $\mathbf{L} \in \mathbb{R}^{m\times n\times m\times n}$, define the matrix $(\mathbf{L} \otimes \mathbf{A})_{ij} = \left(\sum_{kl} L_{ijkl} A_{kl}\right)_{ij}$. Denote the Frobenius matrix inner product by $\langle\cdot,\cdot\rangle_{\mathrm{F}}$. Define the probability simplex $\Delta_J = \{q \in \mathbb{R}^J,\ q_j \geq 0 \text{ for } j = 1,\dots,J,\ \sum_j q_j = 1\}$. Finally, $\boldsymbol{x}_{:i}$ corresponds to the first i time steps of $\boldsymbol{x}$.

2 Dynamic Time Warping for Time Series Alignment

Sakoe and Chiba (1978) consider the problem of aligning two time series $\boldsymbol{x} \in \mathcal{X}^{T_x}$ and $\boldsymbol{y} \in \mathcal{X}^{T_y}$, where potentially $T_x \neq T_y$. This is formalized as

$$\mathrm{DTW}(\boldsymbol{x},\boldsymbol{y}) = \min_{\mathbf{A}\in\mathcal{A}(T_x,T_y)} \langle \mathbf{D}, \mathbf{A}\rangle_{\mathrm{F}} \qquad (1)$$

where $D_{ij} = d_{\mathcal{X}}(x_i, y_j)$ is the pairwise distance matrix. This problem amounts to finding an alignment matrix that minimizes the total alignment cost. The objective (1) can be computed in $O(T_xT_y)$ by leveraging the dynamic programming forward recursion

$$\begin{aligned}&\mathrm{DTW}(\boldsymbol{x}_{:i},\boldsymbol{y}_{:j}) = d_{\mathcal{X}}(x_i,y_j)\\ &+\min(\mathrm{DTW}_{i-1,j},\mathrm{DTW}_{i-1,j-1},\mathrm{DTW}_{i,j-1}),\end{aligned} \qquad (2)$$

where $\mathrm{DTW}_{i,j} = \mathrm{DTW}(\boldsymbol{x}_{:i},\boldsymbol{y}_{:j})$. The optimal alignment matrix $\mathbf{A}^*$ can then be obtained by tracking the optimal path backwards. DTW is a more flexible choice for comparing time series than element-wise Euclidean distances, because it allows one to compare time series of different sampling frequencies due to its ability to "warp" time. In particular, two time series can be close in DTW even if $T_x \neq T_y$. DTW has been used in a number of settings, including time series averaging, clustering (Petitjean and Gançarski, 2012; Schultz and Jain, 2018) and feature extraction (Yi et al., 1998; Kate, 2016).

A limitation of DTW is the discontinuity of its gradient, which can affect the performance of gradient descent algorithms. To address this, Cuturi and Blondel (2017) introduced a soft version of DTW. The minimum in (1) is replaced with a softened version, yielding

$$\mathrm{DTW}_\gamma(\boldsymbol{x},\boldsymbol{y}) = -\gamma \log \sum_{\mathbf{A}\in\mathcal{A}(T_x,T_y)} \exp\left(-\tfrac{1}{\gamma}\langle \mathbf{D},\mathbf{A}\rangle_{\mathrm{F}}\right). \qquad (3)$$

DTW is recovered in the limit $\gamma \to 0$. They also discuss a softened version of the optimal alignment matrix $\mathbf{A}^*$, given by the softened argmin

$$\operatorname*{arg\,min}^{\gamma}_{\mathbf{A}\in\mathcal{A}(T_x,T_y)} \langle \mathbf{D},\mathbf{A}\rangle_{\mathrm{F}} = C_{\boldsymbol{x},\boldsymbol{y}}^{-1} \sum_{\mathbf{A}\in\mathcal{A}(T_x,T_y)} \exp\left(-\tfrac{1}{\gamma}\langle \mathbf{D},\mathbf{A}\rangle_{\mathrm{F}}\right)\mathbf{A}, \quad (4)$$

where $\gamma \geq 0$ is a smoothing parameter and $C_{\boldsymbol{x},\boldsymbol{y}}$ is the normalizing constant of the unnormalized density $P(\mathbf{A}) \propto e^{-\frac{1}{\gamma}\langle \mathbf{D},\mathbf{A}\rangle_{\mathrm{F}}}$. While they consider temporal variability, DTW and soft DTW are not invariant under transformations, such as translations and rotations, which can limit their application to settings where time series are obtained only up to isometric transformations, such as word embeddings. To alleviate this, Vayer et al. (2020) propose

$$\mathrm{DTW\text{-}GI}(\boldsymbol{x},\boldsymbol{y}) = \min_{f\in\mathcal{F}} \mathrm{DTW}(\boldsymbol{x}, f(\boldsymbol{y})), \qquad (5)$$

which gives a distance between time series that is invariant under a set of transformations $\mathcal{F}$, where f is applied elementwise to points of the time series; Vayer et al. (2020) consider orthonormal transformations, such as rotations. In more general settings, this requires one to optimize over a potentially large space of transformations $\mathcal{F}$, which becomes infeasible if $\boldsymbol{x}$ and $\boldsymbol{y}$ are too different.

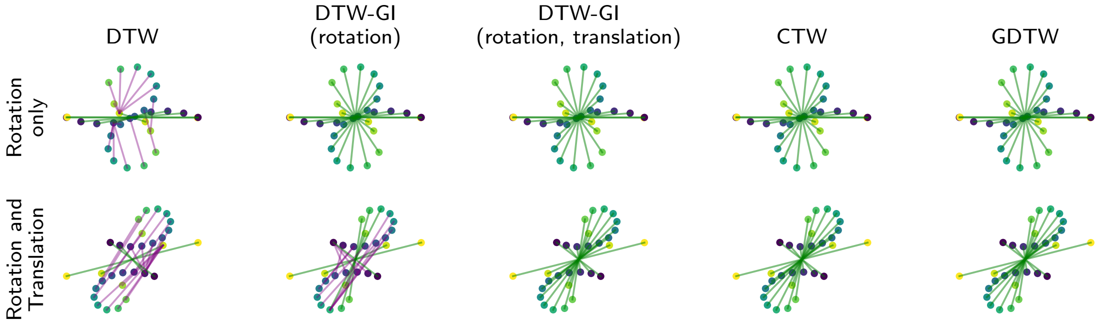


Figure 1: Alignment of time series equivalent up to rotation by 180 degrees (top), and up to rotation and translation (bottom). Node coloring represents time (dark purple: $t = 0$, yellow: $t = T$), and edge coloring represents alignment correctness (respectively green and purple for correct and incorrect matchings). The CTW and DTW-GI (rotation, translation) baselines recover the correct alignment. GDTW also recovers the correct alignments, but without needing to manually specify a cost function or symmetries. DTW-GI (rotation) fails in the translational setting, and DTW fails in both.

Similarly to DTW-GI, Canonical Time Warping (CTW) (Zhou and Torre, 2009) consists of aligning the data temporally via DTW and spatially via canonical correlation analysis (CCA). CTW is defined as

$$\mathrm{CTW}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\substack{\mathbf{W}_x, \mathbf{W}_y \\ \mathbf{V}_x, \mathbf{V}_y}} \|\mathbf{V}_x \boldsymbol{x} \mathbf{W}_x + \mathbf{V}_y \boldsymbol{y} \mathbf{W}_y\|_{\mathrm{F}}^2 \qquad (6)$$

with constraints on matrices $\mathbf{W}_x, \mathbf{W}_y, \mathbf{V}_x, \mathbf{V}_y$, which make CTW invariant to translations, rotations and scaling at optimality. Optimization is performed by alternation on $\mathbf{V}_x, \mathbf{V}_y$ via DTW, and on $\mathbf{W}_x, \mathbf{W}_y$ via CCA. In particular, the former matrices align $\boldsymbol{x}$ and $\boldsymbol{y}$ temporally whilst the latter ones align the time series spatially by projecting the temporally-aligned time series onto a common subspace on which they are maximally correlated. Zhou and De la Torre (2016) generalize CTW to allow for the alignment of multiple time series, and Trigeorgis et al. (2018) allow for non-linear projections. Gong and Medioni (2011) leverage manifold learning to align the time series spatially in conjuction with DTW for temporal alignment.

2.1 Connecting DTW and Optimal Transport

Optimal transport (Peyré and Cuturi, 2019) allows one to compare and average measures in a way that incorporates the geometry of the underlying space on which they are defined. Such approaches can be intuitively connected to DTW by observing that time series are essentially discrete measures equipped with an ordering. This allows one to view the alignment matrices in the DTW objective as analogues of coupling matrices that appear in the Kantorovich formulation of the classical optimal transport problem (Villani, 2008). To formalize this, consider the Wasserstein distance between discrete measures. Let $\mu_x = \sum_{i=1}^m p_i \delta_{x_i}$, $\mu_y = \sum_{i=1}^n q_i \delta_{y_i}$ be discrete probability measures with $\boldsymbol{p} \in \Delta_m, \boldsymbol{q} \in \Delta_n$, and set $D_{ij} = d_{\mathcal{X}}(x_i, x_j)$. Define the Wasserstein distance between discrete measures μ_x and μ_y as

$$\mathrm{W}(\mu_x, \mu_y) = \min_{\mathbf{T} \in \Pi(\boldsymbol{p}, \boldsymbol{q})} \langle \mathbf{D}, \mathbf{T} \rangle_{\mathrm{F}}, \qquad (7)$$

where $\Pi(\boldsymbol{p}, \boldsymbol{q})$ is the set of coupling matrices with marginals $\boldsymbol{p}$ and $\boldsymbol{q}$. Equation (7) clearly resembles (1), and in both cases the objective consists of the minimization of the element-wise dot product between a distance matrix and another matrix, which we term the *plan*. In the DTW case, the plan consists of an alignment matrix, and in the Wasserstein case it consists of a coupling matrix. Moreover, the optimal coupling T_{ij}^* describes the optimal amount of probability mass to move from point x_i to y_j, whilst the optimal alignment A_{ij}^* describes whether or not x_i and y_j are aligned at optimality. While tightly connected, DTW and the Wasserstein distance between time series' support points are still different. For example, if we consider two time series with the same points but reversed ordering, these would be far away under DTW, but equal under Wasserstein.

The Wasserstein distance is limited by the requirement for a sensible ground metric $d_{\mathcal{X}}$ to be defined between samples $x_i \in \mathcal{X}$ and $y_j \in \mathcal{Y}$, which is impossible if there does not exist an explicit correspondence between samples from the compared measures (Solomon et al., 2016). The Wasserstein distance is also not invariant under isometries, such as rotations and translations, and generally leads to a large distance between measures equivalent up to such transformations. To relax these requirements, Mémoli (2011) propose the *Gromov–Wasserstein* (GW) distance between isometry classes of metric-measure triples $(\mathcal{X}, d_{\mathcal{X}}, \mu_x)$ and

$(\mathcal{Y}, d_{\mathcal{Y}}, \mu_y)$. It is defined as

$$\begin{aligned}&\mathrm{GW}(\mu_x, \mu_y)\\ &= \min_{\mathbf{T}\in\Pi(\boldsymbol{p},\boldsymbol{q})} \sum_{ijkl} \mathcal{L}\big(d_{\mathcal{X}}(x_i, x_k), d_{\mathcal{Y}}(y_j, y_l)\big) T_{ij} T_{kl},\end{aligned} \tag{8}$$

where $\mathcal{L}$ is typically squared error loss, and does not rely on a cost or metric to compare x_i with y_j. Instead, GW compares the intra-relational metric geometries of the two measures by comparing the distributions of their pairwise distances. This only requires the definition of metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively, which can be arbitrarily different. GW has been used as a tool for comparing measures on incomparable spaces, notably for training generative models (Bunne et al., 2019), graph matching (Xu et al., 2019b), and graph averaging (Xu et al., 2019a). Vayer et al. (2019) also propose *fused Gromov–Wasserstein* to deal with structured objects such as graphs and time series, which consists of a mixture of Wasserstein distance on the node features (for example, time ordering), and GW on the spatial structure, which illustrates how these concepts can be mixed and matched as needed in the specific use case.

3 Gromov Dynamic Time Warping

Motivated by the connections between DTW and optimal transport described in Sections 2 and 2.1, respectively, we introduce a distance between time series $\boldsymbol{x} \in \mathcal{X}^{T_x}$ and $\boldsymbol{y} \in \mathcal{Y}^{T_y}$ defined on potentially incomparable compact metric spaces. We define the *Gromov dynamic time warping* distance between metric-time-series triples $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})$ as

$$\begin{aligned}&\mathrm{GDTW}(\boldsymbol{x}, \boldsymbol{y})\\ &= \min_{\mathbf{A}\in\mathcal{A}(T_x,T_y)} \sum_{ijkl} \mathcal{L}\big(d_{\mathcal{X}}(x_i, x_k), d_{\mathcal{Y}}(y_j, y_l)\big) A_{ij} A_{kl},\end{aligned} \tag{9}$$

where $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}^+$ is a loss function measuring the alignment of the pairwise distances. The first two elements of the metric-time-series triples are omitted to ease notation. We think of $\mathcal{L}$ as a proxy for measuring the alignment of the time series (e.g., the square error loss $\mathcal{L}(a, b) = (a - b)^2$). Under the optimal alignment, for any two pairs (x_i, y_j) and (x_k, y_l), if x_i is close to x_k then y_j will tend to be close to y_l.

Provided $\mathcal{L}$ is a pre-metric and so induces a Hausdorff topology, GDTW possesses the following properties:

(a) $\mathrm{GDTW}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, and $\mathrm{GDTW}(\boldsymbol{x}, \boldsymbol{x}) = 0$,

(b) $\mathrm{GDTW}(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if there exists an isometry $\phi : \mathcal{X} \to \mathcal{Y}$ such that $\phi(\boldsymbol{x}) = \boldsymbol{y}$,

(c) $\mathrm{GDTW}(\boldsymbol{y}, \boldsymbol{x}) = \mathrm{GDTW}(\boldsymbol{x}, \boldsymbol{y})$ if and only if $\mathcal{L}$ is symmetric.

Algorithm 1 Frank–Wolfe-inspired algorithm for Gromov DTW

```
Initialize $\mathbf{A} \in \mathcal{A}(T_x, T_y)$ arbitrarily, and compute $L_{ijkl} = \mathcal{L}\big(d_{\mathcal{X}}(x_i, x_k), d_{\mathcal{Y}}(y_j, y_l)\big)$.
while iter < max_iter and has not converged do
    Update $\mathbf{A} \leftarrow \arg\min^{\gamma}_{\mathbf{A}'\in\mathcal{A}(T_x,T_y)} \langle \mathbf{L} \otimes \mathbf{A}, \mathbf{A}' \rangle_{\mathrm{F}}$ using (2) if $\gamma = 0$ or (17) if $\gamma > 0$.
end while
return A
```

Mirroring DTW, GDTW does not generally satisfy the triangle inequality. Thus, GDTW is a pre-metric over equivalence classes of $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x})$ triples, up to metric isometry. A formal treatment is given in Appendix A.

Some optimal alignments are given in Figure 1. The original version of DTW-GI (rotationally invariant) fails in the translational case, while its translational extension, obtained by subtracting a bias from both time series, works in both cases—here, invariances have to be manually specified. CTW works in both settings, but invariances are also manually specified by the constraints imposed in the optimization of the learned spatial projections. GDTW recovers the correct alignments in both cases without explicitly specifying the symmetries.

3.1 A Frank–Wolfe-inspired Algorithm

We now present a straightforward and efficient algorithm for computing GDTW. Following ideas proposed in the optimal transport setting for computing the Gromov–Wasserstein distance, one can introduce a 4-dimensional array $L_{ijkl} = \mathcal{L}\big(d_{\mathcal{X}}(x_i, x_k), d_{\mathcal{Y}}(y_j, y_l)\big)$ and express GDTW as

$$\mathrm{GDTW}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\mathbf{A}\in\mathcal{A}(T_x,T_y)} \mathcal{G}_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{A}), \tag{10}$$

$$\mathcal{G}_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{A}) = \langle \mathbf{L} \otimes \mathbf{A}, \mathbf{A} \rangle_{\mathrm{F}}. \tag{11}$$

This expression is similar to the DTW objective in (1), but with a cost function $\mathbf{D}$ that now depends on the alignment matrix $\mathbf{A}$.

The Frank–Wolfe (FW) method is an algorithm for solving constrained optimization problems without requiring projections onto the constraint set. While FW optimization on convex domains has been deeply studied for both convex (Frank and Wolfe, 1956; Jaggi, 2013) and non-convex (Lacoste-Julien, 2016) objectives, FW on non-convex domains is largely unexplored. Inspired by the non-convex Frank–Wolfe algorithm introduced in Balashov et al. (2020), we propose a variant

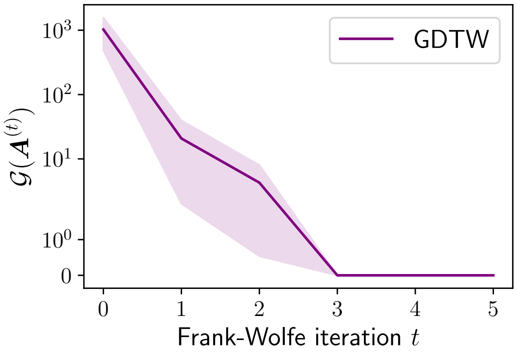


Figure 2: Evolution of the Gromov DTW objective with respect to iteration number for the time series of Figure 1. We plot mean and standard deviation across 10 runs with randomly initialized alignment matrices.

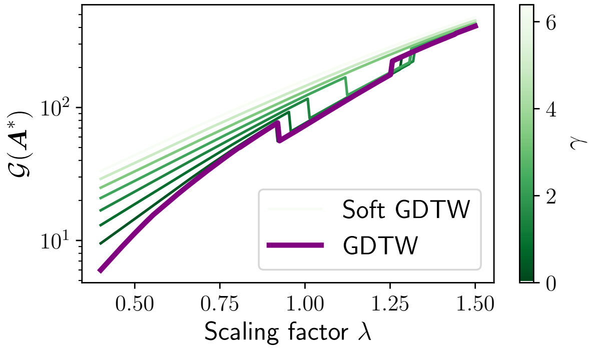


Figure 3: Evolution of (soft) GDTW($\boldsymbol{x}, \boldsymbol{x}_\lambda$), where $\boldsymbol{x}_\lambda$ is obtained by distorting the first $T/2$ points of $\boldsymbol{x}$ by λ. As $\gamma \to \infty$, soft GDTW becomes smoother and the jumps disappear. As $\gamma \to 0$, it converges to GDTW.

that enforces feasibility of proposals by setting the step size to 1. Our algorithm consists of the following steps. First, we (i) solve a linear minimization oracle

$$\mathbf{S}^{(t)} = \underset{\mathbf{A}\in\mathcal{A}(T_x,T_y)}{\arg\min} \left\langle \nabla_{\mathbf{A}} \mathcal{G}_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{A}^{(t)}), \mathbf{A} \right\rangle \tag{12}$$

$$= \underset{\mathbf{A}\in\mathcal{A}(T_x,T_y)}{\arg\min} \left\langle \mathbf{L} \otimes \mathbf{A}^{(t)}, \mathbf{A} \right\rangle, \tag{13}$$

which can be performed exactly in $O(T_x T_y)$ by a DTW iteration, noting that $\mathbf{L} \otimes \mathbf{A}^{(t)}$ can be computed in $O(T_x^2 T_y + T_x T_y^2)$ time in the case $\mathcal{L} = L_2$ (Peyré et al., 2016). Then, we (ii) updates the iterates. For the step size $\eta^{(t)} = 1$, the update is

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta^{(t)}(\mathbf{S}^{(t)} - \mathbf{A}^{(t)})) = \mathbf{S}^{(t)}. \tag{14}$$

Keeping step sizes $\eta^{(t)}$ in $\{0, 1\}$ remediates the non-convexity of the constraint set, as iterates are guaranteed to remain in $\mathcal{A}(T_x, T_y)$ in spite of non-convexity.

In Figure 2, we plot the objective $\mathcal{G}_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{A}^{(k)})$ at each iteration k across various initializations of alignment matrices, for the time series illustrated in the top row of Figure 1. We observe that in this example, the algorithm recovers the optimal alignment with loss value 0 in a handful of iterations and is robust with respect to to initialization.

Due to the discrete nature of alignment matrices in the GDTW objective, providing convergence guarantees is non-trivial. We thus focus on empirical evaluation in Section 5 across various settings (such as barycentric averaging, generative modeling, and imitation learning) to demonstrate that the method works well in practice, and defer convergence analysis to future work. In practice, we terminate Algorithm 1 if it converges, potentially to a limit cycle, or if the number of iterations reaches a fixed threshold. A number of alternative algorithms are possible and could be developed, for instance through solving the inner minimization oracle on the convex hull of $\mathcal{A}$ and projecting the results onto the constraint set—we defer these to future work.

3.2 Gromov DTW as a Loss Function

Gromov DTW can be itself used as a differentiable loss function. Here, we apply the envelope theorem (Carter, 2001; Milgrom and Segal, 2002) to (10) and obtain

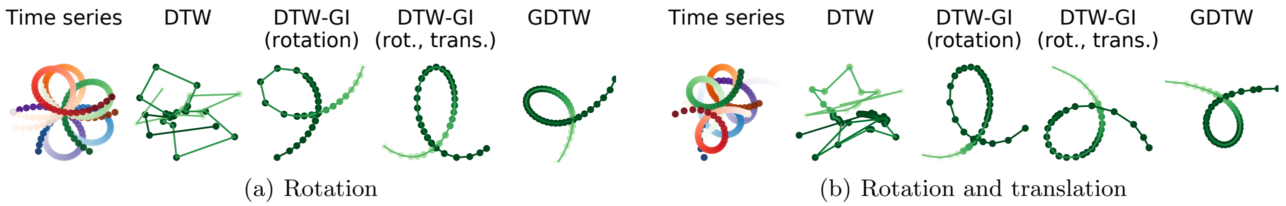
$$\nabla_{\boldsymbol{x},\boldsymbol{y}} \operatorname{GDTW}(\boldsymbol{x}, \boldsymbol{y}) = \nabla_{\boldsymbol{x},\boldsymbol{y}} \langle \mathbf{L}(\boldsymbol{x}, \boldsymbol{y}) \otimes \mathbf{A}^*, \mathbf{A}^* \rangle_{\mathrm{F}}, \tag{15}$$

$$\mathbf{A}^* = \underset{\mathcal{A}(T_x,T_y)}{\arg\min}\, \mathcal{G}_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{A}). \tag{16}$$

Similarly to DTW, GDTW suffers from unpredictability when the time series is close to a change point of the optimal alignment matrix because of the discontinuity of derivatives. To remediate this, we describe how GDTW can be softened analogously to soft DTW, to obtain smoother derivatives. A smoother landscape also helps robustify GDTW with respect to alignment initialization. The algorithm for computing Gromov DTW consists of successive DTW iterations. Following ideas from the Gromov–Wasserstein literature, we replace the DTW operation in the iterations with a softened version, by replacing the argmin by the soft argmin in (4). A priori, it may seem that computing this is significantly more involved. However, Cuturi and Blondel (2017) observe that

$$\underset{\mathbf{A}\in\mathcal{A}(T_x,T_y)}{\arg\min{}^{\gamma}} \langle \mathbf{D}, \mathbf{A} \rangle_{\mathrm{F}} = \nabla_{\mathbf{D}} \operatorname{DTW}_{\gamma}(\mathbf{D}), \tag{17}$$

where $\arg\min^\gamma$ is the softened $\arg\min$ defined in (4). Hence, (4) can be computed by reverse-mode automatic differentiation in quadratic time, and soft GDTW iterations can be performed by plugging in $\mathbf{D} = \mathbf{L} \otimes \mathbf{A}$. We approximate the derivatives of soft GDTW by using the optimal soft alignment matrix and applying (15)



(a) Rotation

(b) Rotation and translation

Figure 4: Barycenters of times series with DTW, DTW-GI, and GDTW. In (a) random rotations are applied to the time series, while in (b) random rotations and translations are applied. DTW fails in both settings, DTW-GI (rotation) fails in the translational setting, while DTW-GI (rotation, translation) and GDTW average sensibly in both as they are invariant to both rotations and translations.

and (16): by the envelope theorem, this approximation becomes exact in the small-γ limit.

In Figure 3, we plot the evolution of GDTW and soft GDTW as one of the 2D time series gets distorted by a factor λ: $\boldsymbol{x}_\lambda = \boldsymbol{x} + (0, \lambda)$. Across a range of λ values GDTW's optimal alignment matrices vary in discrete steps, which leads to discontinuous values, and hence discontinuous gradients, around such $\boldsymbol{x}_\lambda$ values. By contrast, soft GDTW with sufficiently high γ values is qualitatively smooth with respect to λ, which remediates discontinuity of GDTW's gradients.

4 Learning with Gromov DTW as a Loss Function

We now present a range of applications of Gromov DTW, including barycentric averaging, generative modeling and imitation learning.

4.1 Barycenters

To compute barycenters of Gromov DTW (10), we extend the algorithm from Peyré et al. (2016) to the sequential setting. Given time series $\boldsymbol{x}_1, ..., \boldsymbol{x}_J \in \mathcal{X}_1^{T_1}, ..., \mathcal{X}_J^{T_J}$ and weights $\boldsymbol{\alpha} \in \Delta_J$, let $(\mathbf{D}_{\boldsymbol{x}_j})_{mn} = d_{\mathcal{X}_j}(\boldsymbol{x}_j^{(m)}, \boldsymbol{x}_j^{(n)})$. For fixed $T \in \mathbb{N}$ (length of the barycentric time series), the barycenter is defined as any triple $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x})$ satisfying

$$\mathbf{D}^* = \operatorname*{arg\,min}_{\mathbf{D} \in \mathbb{R}^{T \times T}} \sum_{j=1}^{J} \alpha_j \operatorname{GDTW}(\mathbf{D}, \mathbf{D}_{\boldsymbol{x}_j}), \tag{18}$$

$$\mathbf{D}_{mn} = d_{\mathcal{X}}(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(n)}), \ n, m = 1, \dots, T, \tag{19}$$

where, to ease notation, we denote GDTW purely in terms of distance matrices. The barycentric time series can then be reconstructed by applying multidimensional scaling (MDS) (Kruskal and Wish, 1978) to $\mathbf{D}^*$: see Figure 4 for an illustration. We rewrite (18) as

$$\min_{\substack{\mathbf{D} \in \mathbb{R}^{T \times T} \\ \mathbf{A}_1, ..., \mathbf{A}_J \in \mathcal{A}(T_x, T_y)}} \sum_{j=1}^{J} \alpha_j \langle \mathcal{L}(\mathbf{D}, \mathbf{D}_{\boldsymbol{x}_j}) \otimes \mathbf{A}_j, \mathbf{A}_j \rangle_{\mathrm{F}} \tag{20}$$

and solve it by alternating between minimizing over $\mathbf{A}_j$ for $j \in 1, ..., J$ via Algorithm 1, and minimizing over $\mathbf{D}$ for fixed $\mathbf{A}_j$. The latter step admits a closed-form solution given as follows.

Proposition 1. *If $\mathcal{L}$ is squared error loss, the solution to the minimization in (20) for fixed $\mathbf{A}_j$ is*

$$\mathbf{D} = \sum_{j=1}^{J} \frac{\alpha_j \mathbf{A}_j^T \mathbf{D}_{\boldsymbol{x}_j} \mathbf{A}_j}{\sum_{j=1}^{J} \alpha_j (\mathbf{A}_j \mathbf{1})(\mathbf{A}_j \mathbf{1})^T}, \tag{21}$$

where division is performed element-wise, and $\mathbf{1}$ is a vector of ones.

Proof. Appendix A. □

4.2 Generative Modeling

We now use GDTW as an approach for training generative models of time series. Here, we view our dataset of time series $\boldsymbol{x}^1, ..., \boldsymbol{x}^J \in \mathcal{X}_1^{T_1}, ..., \mathcal{X}_J^{T_J}$ as a discrete measure $\mu = \frac{1}{J} \sum_{j=1}^{J} \delta_{\boldsymbol{x}^j}$. We define a generative model $\mu_\theta = G_{\theta\#}\nu$, where ν is a latent measure, such as an isotropic Gaussian, $G_\theta : \mathcal{Z} \to \mathcal{X}^T$ is a neural network and $G_{\theta\#}\nu$ is the pushforward measure. By nature of Gromov DTW, the generated time series do not have to live in the same space as the data. In particular, this allows us to specify the length of the time series we wish to generate. We train the model μ_θ by minimizing the entropic Wasserstein distance W_ε (Cuturi, 2013) between μ and μ_θ. For the ground cost d of W_ε, we use DTW_γ and GDTW_γ. For GDTW_γ, the objective is

$$\begin{aligned} &\min_{\theta \in \Theta} \mathrm{W}_\varepsilon(\mu, \mu_\theta) \\ &= \min_{\pi \in \Pi(\mu, \mu_\theta)} \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \pi} \mathrm{GDTW}_\gamma(\boldsymbol{x}, \boldsymbol{y}) - \varepsilon H(\pi), \end{aligned} \tag{22}$$

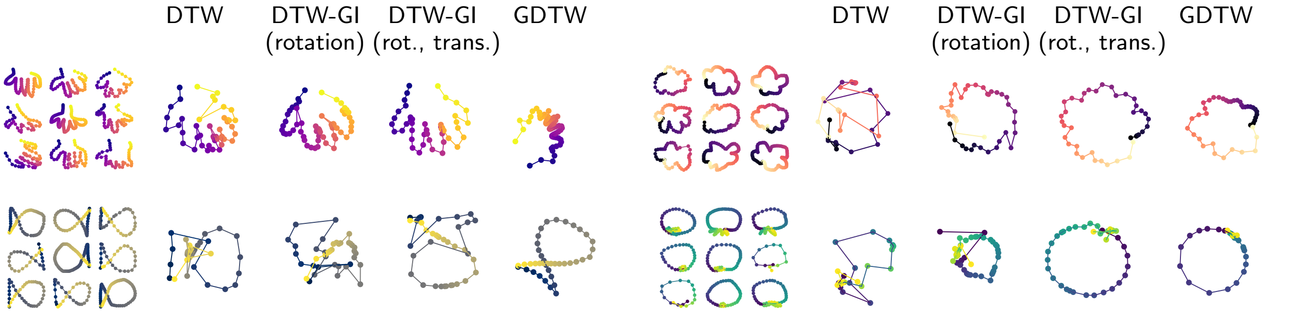


Figure 5: Barycenters computed on the QuickDraw dataset using DTW, DTW-GI and GDTW, and sample data points from four different classes (*hands*, *clouds*, *fishes*, *blueberries*). We observe that only GDTW barycenters are meaningful across all datasets, and hence that GDTW better captures the geometric shape of the time series.

where H is the entropic regularization term. Following Genevay et al. (2018), it is also possible to use the debiased analog of (22). $\mathrm{W}_\varepsilon(\mu, \mu_\theta)$ is computed efficiently using the Sinkhorn algorithm (Sinkhorn, 1974; Cuturi, 2013), and θ is minimized by gradient descent. This approach extends the Sinkhorn GAN by Genevay et al. (2018) and the GWGAN by Bunne et al. (2019) to sequential data.

4.3 Imitation Learning

We consider an imitation learning setting in which an agent needs to solve a task given the demonstration of an expert. We assume the agent has access to the true transition function $\mathcal{T}$ over the agent's state-space $\mathcal{X}$, and define a state trajectory as a time series $\boldsymbol{x} \in \mathcal{X}^{T_x}$. An expert state trajectory $\boldsymbol{y}_{\text{exp}} \in \mathcal{Y}^{T_y}$ solving a specific task, such as traversing a maze, is given. The goal is to train the agent's parametrized policy $\pi_\theta : \mathcal{X} \to \mathcal{A}$ to solve the given task by imitating the expert's behavior, where $\mathcal{A}$ is the action space. To find this policy, the agent uses the model of the environment to predict state trajectories $\boldsymbol{x}_\theta$ under the current policy π_θ, compares these predictions with the expert's trajectory $\boldsymbol{y}_{\text{exp}}$, and then optimizes the controller parameters θ to minimize the distance between predicted agent trajectory and observed expert trajectory. Using GDTW, our objective is

$$\min_\theta \mathrm{GDTW}_\gamma(\boldsymbol{y}_{\text{exp}}, \boldsymbol{x}_\theta). \tag{23}$$

The flexibility of GDTW allows for expert trajectories defined in pixel space $\mathcal{Y} = \mathbb{R}^{32\times 32}$, while the agent lives in $\mathcal{X} = \mathbb{R}^2$. Rollouts obtained with π_θ mimic the expert's trajectory up to isometry. For comparison, instead of (23), we also consider DTW. The aim is to learn the same trajectory in the same space as the expert. DTW, in contrast with GDTW, requires $\mathcal{X} = \mathcal{Y}$, and the starting positions for the agent and expert to be close. From a reinforcement learning perspective, the use of GDTW in (23) can be interpreted as a value estimate and gradient-based policy learning as taking estimated value gradients (Fairbank and Alonso, 2012; Heess et al., 2015).

5 Experiments

We assess the effectiveness of our proposals in settings in which (i) time series live in comparable spaces and where previous approaches apply, (ii) the spaces are incomparable.

Baselines. Throughout the experiments, we compare GDTW_γ to, in settings in which they apply, DTW_γ (Sakoe and Chiba, 1978; Cuturi and Blondel, 2017) its respectively rotationally-invariant and translationally-rotationally-invariant extensions DTW-GI (rotation),

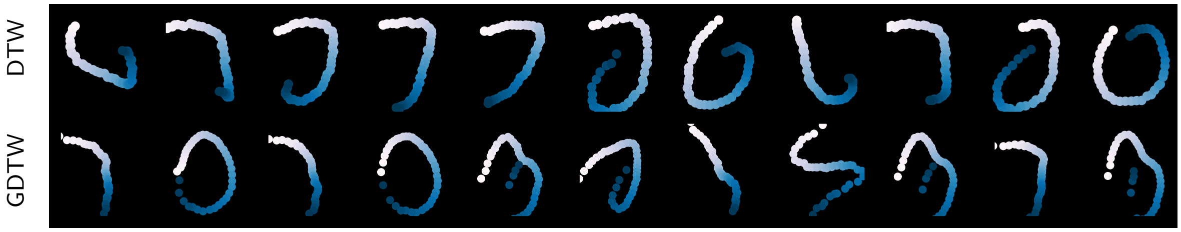
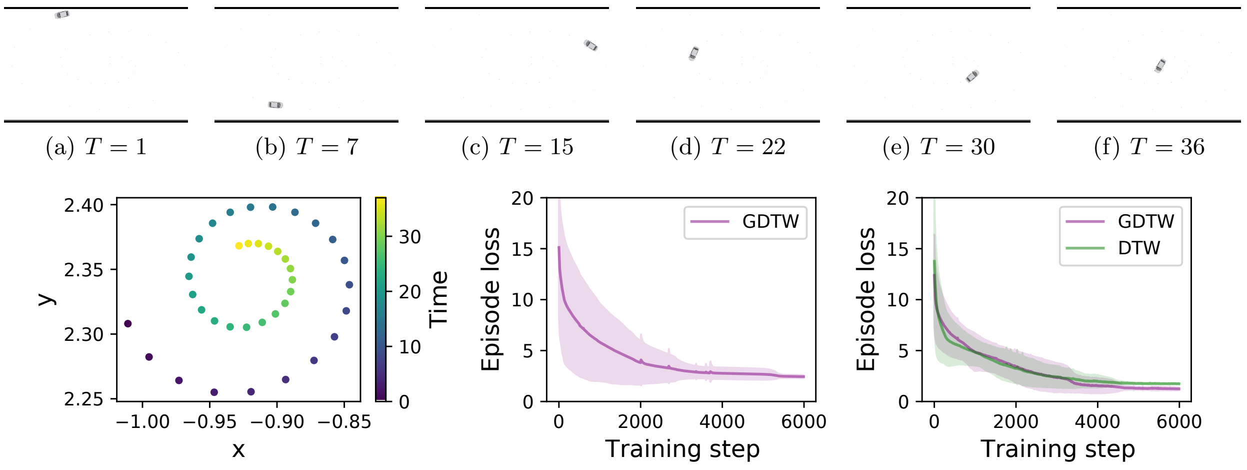


Figure 6: Samples generated by the time series GAN trained on Sequential MNIST, with DTW_γ and GDTW_γ, respectively, used as ground costs.



(g) Rollout of the learned policy (h) Loss: video trajectory (i) Loss: 2D expert trajectory

Figure 7: (a)–(f): Snapshot of an expert trajectory (sequence of pixel images); (g): policy of an agent in $\mathbb{R}^2$ learned by imitation learning given video demonstrations; (h): log-episodic loss per training step in the video/2D setting; (i) in the 2D/2D setting (averaged across 20 seeds, with standard deviations.

DTW-GI (rotation, translation) (Vayer et al., 2020), and canonical time warping (Zhou and Torre, 2009).

5.1 Alignment

We first evaluate GDTW on alignment tasks. We consider two settings in which $\boldsymbol{y}$ is obtained by applying to $\boldsymbol{x}$ (i) a rotation, and (ii) a translation followed by a rotation. In Figure 1, we see that GDTW recovers the right alignment in both settings, while DTW-GI with rotation only works in the rotational setting—this can be seen in the top row of Figure 1. DTW-GI with rotation and translation and CTW work in both settings, while ordinary DTW fails in both. We emphasize that CTW and DTW-GI variants are made invariant to the symmetries by explicitly optimizing manually specified spatial projections, whilst GDTW works in both settings without needing anything to be specified, as GDTW is invariant to symmetries by construction. Further experiments with soft DTW and GDTW are given in Appendix B.

5.2 Barycenter Computation

We investigate barycentric averaging of GDTW, on both toy data and the QuickDraw[1] dataset. We compare Gromov DTW to DTW and DTW-GI variants, where barycenters from the latter two methods are computed using DTW barycentric averaging (Petitjean and Gançarski, 2012).

Toy data. In Figure 4, we see that in comparable settings DTW barycenters fail if time series are rotated or translated. DTW-GI with rotation is robust to rotation, but fails when applying both rotations and translations, because the translational symmetry is not manually specified. By contrast, GDTW is robust to both, and leads to meaningful barycenters in all of the given settings.

QuickDraw dataset. The QuickDraw dataset consists of time series of drawings in $\mathbb{R}^2$, belonging to 345 categories. Among those categories, we selected *hands*, *clouds*, *fishes*, and *blueberries*. To address high variability in classes, we selected input data following a preprocessing routine described in Appendix B. A sample of the data sets, together with barycenters computed with DTW, DTW-GI, and GDTW is displayed in Figure 5. DTW and DTW-GI with rotation fail to reproduce the shape of the inputs for most classes. DTW-GI with rotation and translation outperforms DTW-GI with rotation, but fails on the *fish* class, while GDTW provides meaningful barycenters across the range of examples. GDTW is thus more robust in recovering the geometric shape of the time series, whilst DTW variants are sensitive to isometries.

5.3 Generative Modeling

We evaluate the generative modeling proposal of Section 4.2, and analyze the behavior of the learned model when using DTW and GDTW. Here, we consider the

[1] QuickDraw can be found at HTTPS://QUICKDRAW.WITHGOOGLE.COM/.
[2] Sequential MNIST can be found at HTTPS://GITHUB.COM/EDWIN-DE-JONG/MNIST-DIGITS-STROKE-SEQUENCE-DATA.

sequential-MNIST dataset,[2] which consists of time series of digits in $\mathbb{R}^2$ being drawn, and where each time step corresponds to a stroke. In Figure 6, we see that samples using GDTW as ground cost (22) are of a significantly higher quality than samples using DTW. This can be explained by the variability in the data set: slight translations significantly affect DTW, but not GDTW. Note that the GDTW samples are rotated and reflected, since GDTW only produces learned samples up to metric isometries.

5.4 Imitation Learning

We now apply Gromov DTW to the imitation learning setting of Section 4.3. Here, we are given an expert trajectory $\boldsymbol{y}_{\text{exp}}$, and our goal is to find a policy π_θ, such that the agent's simulated trajectory $\boldsymbol{x}_\theta$ mimics $\boldsymbol{y}_{\text{exp}}$. We consider maze navigation tasks in two settings: (i) both expert trajectories and the agent's domain are $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$ and (ii) expert trajectories consist of a video sequence of 32×32 images, giving $\mathcal{Y} = \mathbb{R}^{32\times32}$, whilst the agent's domain is $\mathcal{X} = \mathbb{R}^2$. In the first setting, DTW and GDTW apply, whilst in the second setting only GDTW can be used. Figure 7(i) displays the loss (23), which is the GDTW distance to the given trajectory, obtained by learning with GDTW and DTW in (i) averaged across 20 seeds. We see that in this fully-comparable setting, GDTW and DTW recover the spiral trajectory provided by the expert.

Finally, we consider a setting in which an agent living in $\mathbb{R}^2$ is provided with an expert trajectory $\boldsymbol{y}_{\text{exp}}$ consisting of a video of a car driving through a spiral, illustrated in Figures 7(a)–7(f) (prior to down-scaling the images). Here, the state-space of the agent, $\mathcal{X} = \mathbb{R}^2$, differs from the state-space of the expert, $\mathcal{Y} = \mathbb{R}^{32\times32}$. The cost on image space $d_{\mathcal{Y}}$ is the 2-Wasserstein distance, with images interpreted as densities on a grid. The cost on the Euclidean space $d_{\mathcal{X}}$ is the Euclidean distance. Figure 7(g) shows the agent's trajectory under the learned policy π_θ, and Figure 7(h) shows the loss (23) against the number of training steps. Using GDTW, the agent successfully learns to solve the task despite never having access to trajectories in the space of interest.

Conclusion

We propose Gromov DTW, a distance between time series living on potentially incomparable spaces. GDTW compares intra-relational geometries of the time series, alleviating the need for a ground metric to be defined on potentially incomparable spaces. Moreover, GDTW is invariant under isometries by nature, which contributes to its versatility and is an important inductive bias for generalization. We hope these contributions enable use of time series alignment in novel settings.

Acknowledgments

We are grateful to K. S. Sesh Kumar for ideas on the Frank–Wolfe algorithm. SC was supported by the Engineering and Physical Sciences Research Council (grant number EP/S021566/1).

References

M. V. Balashov, B. T. Polyak, and A. A. Tremba. Gradient Projection and Conditional Gradient Methods for Constrained Nonconvex Minimization. *Numerical Functional Analysis and Optimization*, 41(7):822–849, 2020. Cited on page 4.

C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka. Learning Generative Models Across Incomparable Spaces. In *ICML*, 2019. Cited on pages 4, 7.

M. Carter. *Foundations of Mathematical Economics*. MIT Press. 2001. Cited on page 5.

M. Cuturi. Fast Global Alignment Kernels. In *ICML*, 2011. Cited on page 1.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NeurIPS*. 2013. Cited on pages 6, 7.

M. Cuturi and M. Blondel. Soft-DTW: A Differentiable Loss Function for Time-Series. In *ICML*, 2017. Cited on pages 1, 2, 5, 7.

M. Fairbank and E. Alonso. Value-Gradient Learning. In *IJCNN*, 2012. Cited on page 7.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré. Interpolating Between Optimal Transport and MMD Using Sinkhorn Divergences. In *AISTATS*, 2019. Cited on page 13.

M. Frank and P. Wolfe. An Algorithm for Quadratic Programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956. Cited on page 4.

A. Genevay, G. Peyre, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *AISTATS*, 2018. Cited on page 7.

D. Gong and G. Medioni. Dynamic Manifold Warping for View Invariant Action Recognition. In *ICCV*, 2011. Cited on page 3.

N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa. Learning Continuous Control Policies by Stochastic Value Gradients. In *NeurIPS*, 2015. Cited on page 7.

M. Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *ICML*, 2013. Cited on page 4.

L. V. Kantorovich. On the Translocation of Masses. *Journal of Mathematical Sciences*, 133(4):1–4, 1958. Cited on page 2.

R. J. Kate. Using Dynamic Time Warping Distances as Features for Improved Time Series Classification. *Data Mining and Knowledge Discovery*, 30(2):283–312, 2016. Cited on page 2.

J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978. Cited on page 6.

S. Lacoste-Julien. Convergence Rate of Frank-Wolfe for Non-Convex Objectives. *arXiv:1607.00345*, 2016. Cited on page 4.

D. Lemire. Faster Retrieval with a Two-Pass Dynamic-Time-Warping Lower Bound. *Pattern Recognition*, 42:2169–2180, 2009. Cited on pages 11, 12.

F. Mémoli. Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. Cited on pages 2, 3.

P. Milgrom and I. Segal. Envelope Theorems for Arbitrary Choice Sets. *Econometrica*, 70:583–601, 2002. Cited on page 5.

F. Petitjean and P. Gançarski. Summarizing a Set of Time Series by Averaging: From Steiner Sequence to Compact Multiple Alignment. *Theoretical Computer Science*, 414:76–91, 2012. Cited on pages 2, 8.

G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. Cited on page 3.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *ICML*, 2016. Cited on pages 5, 6.

H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *ICASSP*, 1978. Cited on pages 1, 2, 7.

D. Schultz and B. Jain. Nonsmooth Analysis and Subgradient Methods for Averaging in Dynamic Time Warping Spaces. *Pattern Recognition*, 74:340–358, 2018. Cited on page 2.

R. Sinkhorn. Diagonal Equivalence to Matrices with Prescribed Row and Column Sums. In *Proceedings of the American Mathematical Society*, 1974. Cited on page 7.

J. Solomon, G. Peyre, V. G. Kim, and S. Sra. Entropic Metric Alignment for Correspondence Problems. *SIGGRAPH*, 2016. Cited on page 3.

G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. Deep Canonical Time Warping for Simultaneous Alignment and Representation Learning of Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1128–1138, 2018. Cited on page 3.

T. Vayer, L. Chapel, N. Courty, R. Flamary, Y. Soullard, and R. Tavenard. Time Series Alignment with Global Invariances. *arXiv:2002.03848*, 2020. Cited on pages 1, 2, 8, 13.

T. Vayer, N. Courty, R. Tavenard, C. Laetitia, and R. Flamary. Optimal Transport for Structured Data with Application on Graphs. In *ICML*, 2019. Cited on page 4.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science, 2008. Cited on pages 3, 11.

H. Xu, D. Luo, and L. Carin. Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching. In *NeurIPS*, 2019. Cited on page 4.

H. Xu, D. Luo, H. Zha, and L. C. Duke. Gromov-Wasserstein Learning for Graph Matching and Node Embedding. In *ICML*, 2019. Cited on page 4.

B.-K. Yi, H. V. Jagadish, and C. Faloutsos. Efficient Retrieval of Similar Time Sequences Under Time Warping. In *ICDE*, 1998. Cited on page 2.

F. Zhou and F. De la Torre. Generalized Canonical Time Warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):279–294, 2016. Cited on page 3.

F. Zhou and F. Torre. Canonical Time Warping for Alignment of Human Behavior. In *NeurIPS*. 2009. Cited on pages 1, 3, 8.

A Theory

Metric Properties

Here we develop the theory of Gromov dynamic time warping distances. We begin by introducing the necessary preliminaries.

Definition 2 (Time series)**.** *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a compact metric space, and let $I_{\mathcal{X}} = \{1, 2, .., T_{\mathcal{X}}\} \subset \mathbb{N}$. We call a finite sequence $\boldsymbol{x} : I_{\mathcal{X}} \to \mathcal{X}$ a* TIME SERIES*. Let X be the space of all time series.*

Definition 3. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be time series. Define a pre-metric $D : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we call the* COST*. Define the $m \times n$* COST MATRIX *$\mathbf{D} \in \mathbb{R}^{m\times n}$ by $D_{ij} = D(x_i, y_j)$.*

Definition 4. *We say that a binary matrix $\mathbf{A}$ is an* ALIGNMENT MATRIX *if $A_{11} = 1$, $A_{mn} = 1$, and $A_{ij} = 1$ implies exactly one of $A_{i-1,j} = 1$, $A_{i,j-1} = 1$, and $A_{i-1,j-1} = 1$ holds. Let*

$$\mathcal{A} = \big\{\mathbf{A} \in \{0, 1\}^{m\times n} : \mathbf{A} \text{ is an alignment matrix}\big\} \tag{24}$$

be the set of ALIGNMENT MATRICES.

Definition 5 (Dynamic Time Warping)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be time series. Define the* DYNAMIC TIME WARPING *distance by*

$$\mathrm{DTW}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\mathbf{A}\in\mathcal{A}} \langle \mathbf{D}, \mathbf{A}\rangle_{\mathrm{F}}, \tag{25}$$

where $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ is the Frobenius norm over real matrices.

Proposition 6. *If D is a pre-metric, then $\mathrm{DTW} : X \times X \to \mathbb{R}$ is a pre-metric on the space of time series. If we take $c = d_{\mathcal{X}}$, then $\mathrm{DTW} : X \times X \to \mathbb{R}$ is a symmetric pre-metric on X.*

Proof. Lemire (2009). □

A pre-metric induces a Hausdorff topology on the set it is defined over, and so is suitable for many purposes that ordinary metrics are used for. To proceed along the path suggested by Gromov-Hausdorff and Gromov–Wasserstein distances over metric-measure spaces, we need to define the time series analog.

Definition 7. *Define a* METRIC SPACE EQUIPPED WITH A TIME SERIES *to be a triple $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x})$.*

Definition 8. *Let $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})$ be metric spaces equipped with time series. Define $X|_{\boldsymbol{x}} = \{x \in X : x \in \operatorname{img} \boldsymbol{x}\}$, and $Y|_{\boldsymbol{y}}$ similarly, and equip both sets with their respective subset metrics. We say that $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})$ are* ISOMORPHIC *if there is a metric isometry $\phi : X|_{\boldsymbol{x}} \to Y|_{\boldsymbol{y}}$ such that $\phi(\widehat{x}_i) = \widehat{y}_i$, where $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$ denote $\boldsymbol{x}$ and $\boldsymbol{y}$ with consecutive repeated elements removed.*

At this stage it is not clear whether or not the class of all such triples under isometry forms a set, or is instead a proper class. To avoid set-theoretic complications, we need the following technical result.

Result 9. *The class of all isometry classes of compact metric spaces is a set.*

Proof. Villani (2008, ch. 27, p. 746). □

It follows immediately that the class of all metric spaces equipped with time series is a set, provided that identification by isometry extends to the time series. We are now ready to define GDTW.

Definition 10. *Let $\mathcal{L}$ be a pre-metric on $\mathbb{R}^+$, and define $\mathcal{L} \in \mathbb{R}^{m\times n\times m\times n}$ by*

$$\mathcal{L}_{ijkl} = \mathcal{L}\big(d_{\mathcal{X}}(x_i, x_k), d_{\mathcal{Y}}(y_j, y_l)\big). \tag{26}$$

Define the GROMOV DYNAMIC TIME WARPING *distance by*

$$\mathrm{GDTW}\big((\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x}), (\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})\big) = \min_{\mathbf{A}\in\mathcal{A}} \langle \mathcal{L} \otimes \mathbf{A}, \mathbf{A}\rangle_{\mathrm{F}}, \tag{27}$$

where $(\mathcal{L} \otimes \mathbf{A})_{ij} = \sum_{kl} L_{ijkl} A_{kl}$.

Proposition 11. GDTW *is a pre-metric on the set of all metric spaces equipped with time series up to isometry.*

Proof. We check the conditions. Non-negativity is immediate by definition. It also follows immediately that $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x}) \cong (\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})$ implies $\operatorname{GDTW}\big((\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x}), (\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})\big) = 0$. We thus need to prove that $\operatorname{GDTW}\big((\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x}), (\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})\big) = 0$ implies $(\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x}) \cong (\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})$. By hypothesis, we have

$$\operatorname{GDTW}\big((\mathcal{X}, d_{\mathcal{X}}, \boldsymbol{x}), (\mathcal{Y}, d_{\mathcal{Y}}, \boldsymbol{y})\big) = \sum_{ijkl} A_{ij} \mathcal{L}_{ijkl} A_{kl} = \sum_{\substack{A_{ij}=1 \\ A_{kl}=1}} \mathcal{L}_{ijkl}, \tag{28}$$

where all elements of the last sum are non-zero. Suppose without loss of generality that $\boldsymbol{x}$ and $\boldsymbol{y}$ contain no duplicate elements. We argue inductively that optimal $\mathbf{A}$ is the identity matrix.

1. First, note that $A_{11} = 1$ by definition of $\mathbf{A}$.

2. Now, consider A_{21}. If we suppose $A_{21} = 1$, then we must have $\mathcal{L}_{2111} = 0$, and hence $d_{\mathcal{X}}(x_2, x_1) = d_{\mathcal{Y}}(y_1, y_1) = 0$. But then $x_2 = x_1$, contradicting the assumption there are no duplicates. Hence, $A_{21} = 0$.

3. By mirroring the above argument, $A_{12} = 0$. Hence, by definition of $\mathbf{A}$, the only remaining possibility is $A_{22} = 1$. Inductively, we conclude $A_{ii} = 1$ for all i, and $A_{ij} = 0$ for $i \neq j$.

4. Finally, since the lower-right corner of $\mathbf{A}$ has to also be equal to one by definition, it follows that $\mathbf{A}$ is the square identity matrix.

Hence $A_{ij} = 1$ and $A_{kl} = 1$ if and only if $i = j$ and $k = l$. Plugging this into the previous equality yields $d_{\mathcal{X}}(x_i, x_k) = d_{\mathcal{Y}}(y_i, y_k)$ for all i, k, which together with diagonal $\mathbf{A}$ gives the isomorphism. Finally, to see that lack of duplicates truly is assumed without loss of generality, note that if there are duplicates in $\boldsymbol{x}$ and $\boldsymbol{y}$, then we apply the above argument to $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$ of Definition 8, which no longer contain duplicates. The claim follows. □

One can easily see that GDTW will be symmetric if L is symmetric. Since DTW itself doesn't satisfy a triangle inequality (Lemire, 2009), GDTW won't satisfy it either.

Barycenter Computation

Proposition 12. *If $\mathcal{L}$ is a square error loss, the solution to the minimization in (20) for fixed $\mathbf{A}_j$ is*

$$\mathbf{D} = \sum_{j=1}^{J} \alpha_j \mathbf{A}_j^T \mathbf{D}_{\boldsymbol{x}_j} \mathbf{A}_j \Big/ \sum_{j=1}^{J} \alpha_j (\mathbf{A}_j \mathbf{1})(\mathbf{A}_j \mathbf{1})^T, \tag{29}$$

where division $\cdot/\cdot$ is performed element-wise, and $\mathbf{1}$ is a vector of ones.

Proof. If $\mathcal{L}$ is square error loss, then (20) can be written as

$$\min_{\mathbf{D}} \sum_{j=1}^{J} \alpha_j \Big\langle \mathbf{D} \odot \mathbf{D} \mathbf{A}_j \mathbf{1} \mathbf{1}^T + \mathbf{1} \mathbf{1}^T \mathbf{A}_j \mathbf{D}_{\boldsymbol{x}_j} \odot \mathbf{D}_{\boldsymbol{x}_j} - 2 \mathbf{D} \mathbf{A}_j \mathbf{D}_{\boldsymbol{x}_j}^T, \mathbf{A}_j \Big\rangle_{\mathrm{F}}, \tag{30}$$

where $\odot$ is element-wise matrix multiplication. Differentiating the objective with respect to $\mathbf{D}$ and setting it equal to 0, we get

$$\mathbf{D} \odot \left(\sum_{j=1}^{J} \alpha_j (\mathbf{A}_j \mathbf{1})(\mathbf{1}^T \mathbf{A}_j^T) \right) = \sum_{j} \alpha_j \mathbf{A}_j^T \mathbf{D}_{\boldsymbol{x}_j} \mathbf{A}_j, \tag{31}$$

which, dividing both sides element-wise, gives the result. □

B Experimental Details

Alignments

In Figures 9–12, we provide further alignment experiments. Note that in this extra set of experiments, we consider the only rotationally invariant proposal of Vayer et al. (2020). Here, we set the entropic term γ to 1 for soft alignments, and we use normalized distance matrices. We observe that GDTW and soft GDTW are robust to scaling, rotations and translations, whilst DTW and soft DTW are sensitive to rotations and translations. Finally, DTW-GI (rotation) is robust to rotations, but sensitive to translations, which further corroborates the observations from Figure 1.

Barycenters

In this experiment, we perform barycenters of 30 elements of 4 quickdraw classes with respect to DTW, DTW-GI and GDTW.

Data selection and pre-processing. The classes considered in the experiment are *fish*, *blueberries*, *clouds* and *hands*. The variability in each class of QuickDraw is extremely high: we created datasets of 30 elements such that it is straightforward to recognize to which category the element belongs to, such that the element is drawn with a single stroke and such that it has a common style. The full datasets are displayed in Figure 8. Before running the algorithms, we rescale the data, applying the transformation $\boldsymbol{x} \mapsto (\boldsymbol{x} - \min(\boldsymbol{x}))/\max(\boldsymbol{x})$ to each data point. Finally, we down-sample the length of the time series reducing it by 1/3 for *hands* and 1/2 for *fish*, *clouds* and *blueberries*.

Algorithms. For GDTW barycenters, we apply the algorithm of Section 4.1, using the entropy regularized version of GDTW with $\gamma = 1$. For DTW and DTW-GI, we use standard DBA procedures. For both algorithms, we set the barycentric length to 60 for *fish* and *hands* and 40 for *clouds* and *blueberries*. We set the maximum number of FW iterations for GDTW to 25, and the number of DTW-GI iterations to 30.

Generative Modeling

In this experiment, we use the Sinkhorn divergence objective. We use a latent dimension of 15, and the generator is a 4-layer MLP with 1000 neurons per layers. The length of the generated time series is set to $T = 40$, and the dimension of the space is $p = 2$, thus the MLP's output dimension is $T \times p = 80$. We set the batch size to 25. We use the ADAM optimizer, with $\boldsymbol{\beta} = (0.5, 0.99)$, and the learning rate set to 5×10^{-5}. We set $\gamma = 1$, and the maximum number of iterations in the GDTW computation to 10. We use the sequential MNIST dataset[3] and normalize the data, which is a time series in $\mathbb{R}^2$, into the unit square.

Imitation Learning

In this experiment, we use a two-layer MLP policy, with input dimension of $\dim(\mathcal{X})$, a hidden dimension of 64, and an output dimension of 2. The learning rate is set to 5×10^{-5}, and we use the ADAM optimizer with $\boldsymbol{\beta} = (0.5, 0.99)$. In the video/2D experiment,[4] the ground cost for the video is entropic 2-Wasserstein distance, computed efficiently using GeomLoss (Feydy et al., 2019), and the ground cost on the 2D space is squared error loss. We plot mean scores along with standard deviations (across 20 random seeds).

[3] Sequential MNIST can be found at https://github.com/edwin-de-jong/mnist-digits-stroke-sequence-data.

[4] The video was generated using https://github.com/gezichtshaar/PyRaceGame.

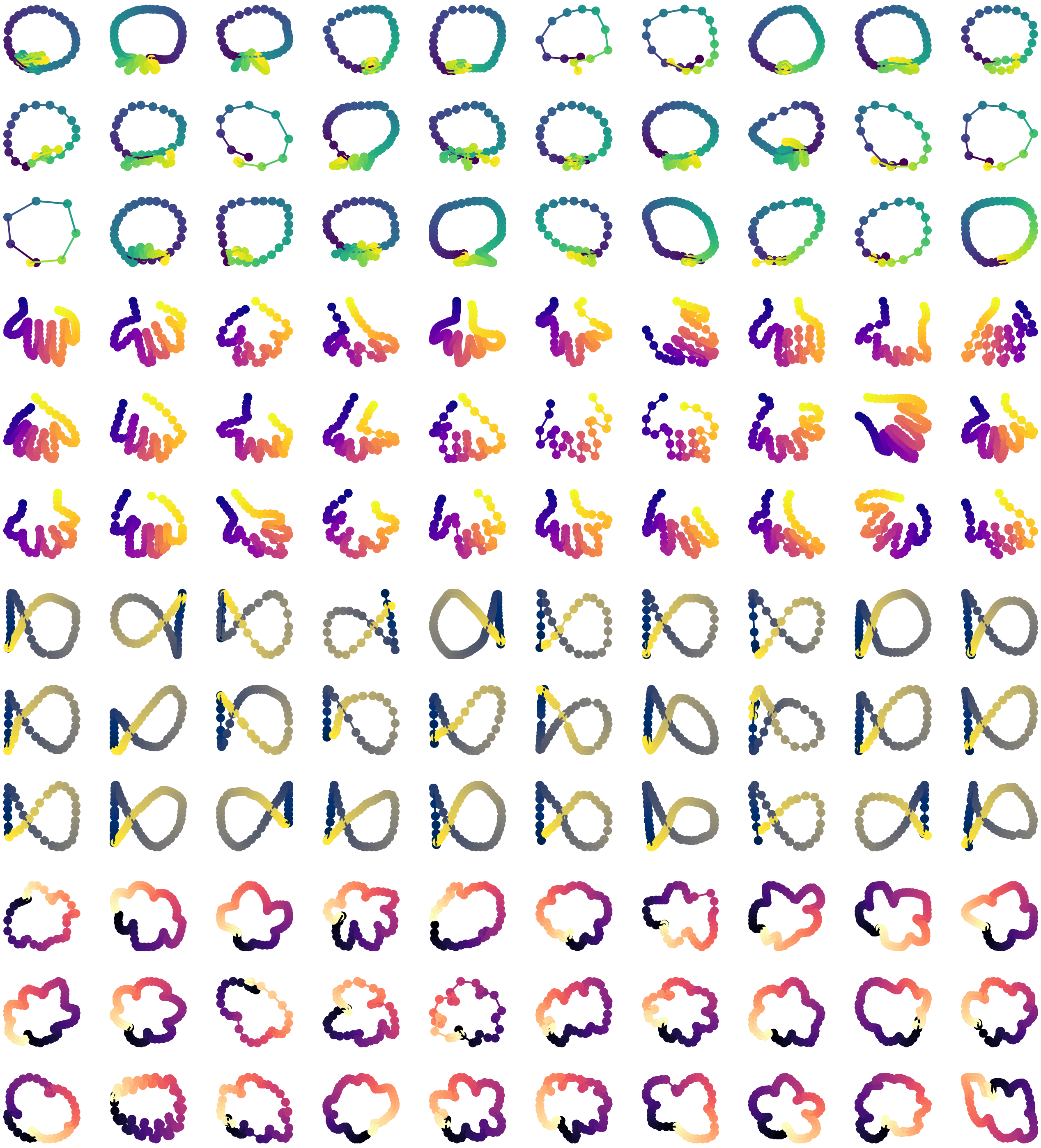


Figure 8: Quickdraw datasets, with classes *blueberries*, *hands*, *fishes*, *clouds*.

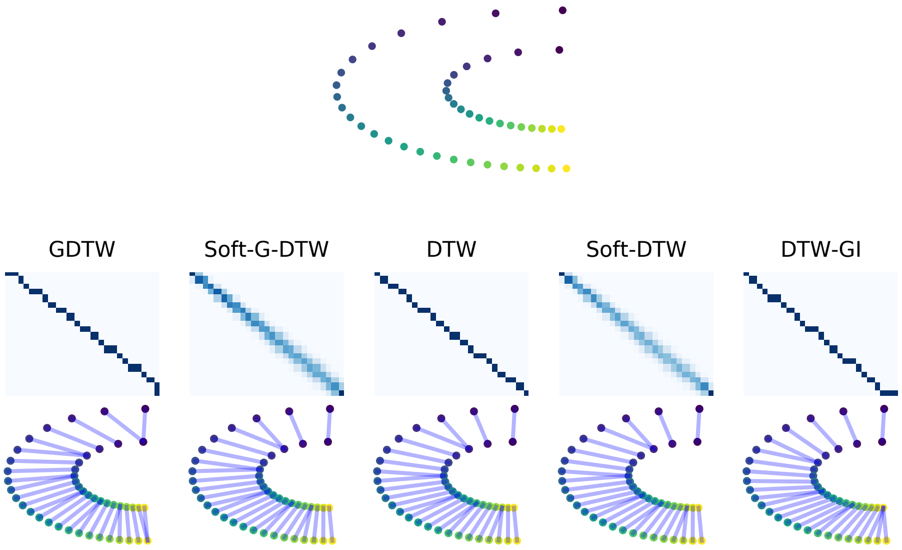


Figure 9: Two time series (top) along with alignment matrices (middle) and alignments with different approaches. In this example, all methods provide a sensible alignment because the time series are on the same axis of rotation and close in the ground space.

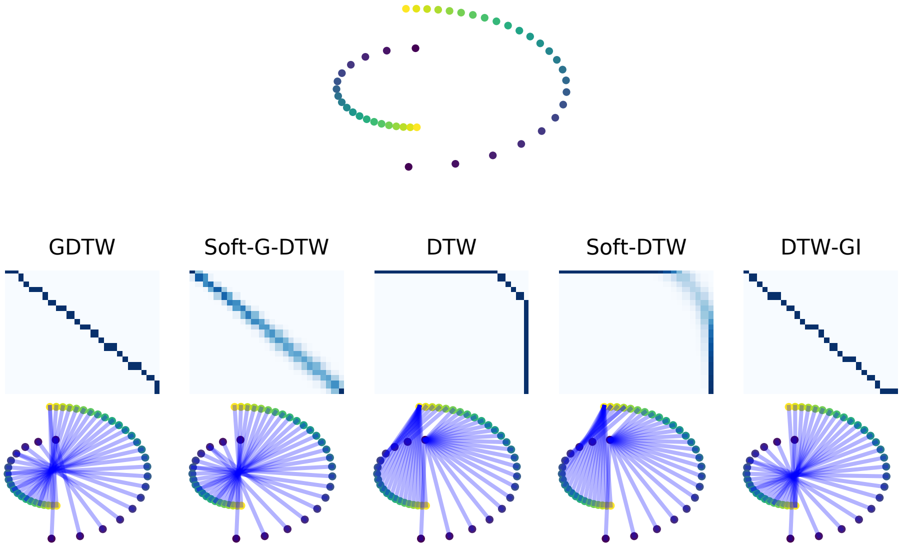


Figure 10: Two time series (top), alignment matrices (middle) and alignments with different approaches. In this example, the time series are not on the same rotation axis which makes DTW variants fail, whilst GDTW and DTW-GI (rotation) provide good alignments due to rotational invariance.

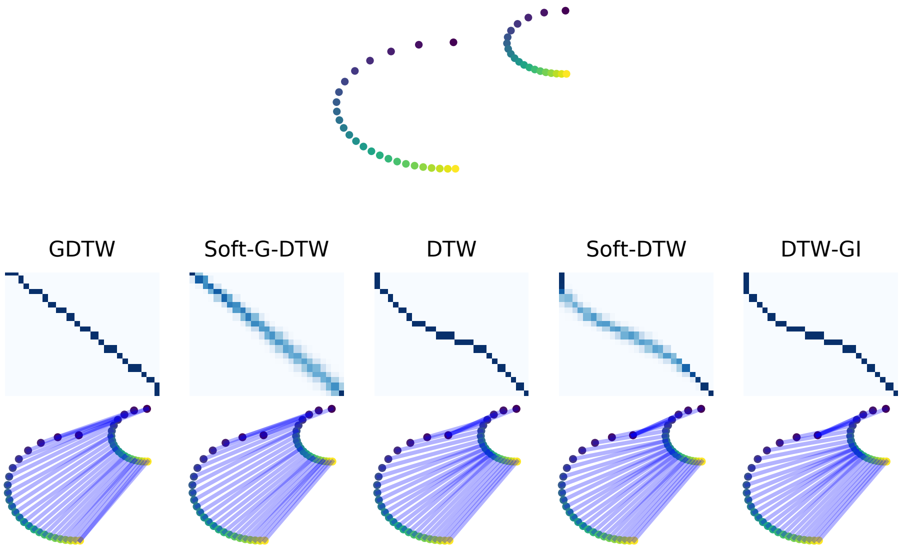


Figure 11: Two time series (top) along with alignment matrices (middle) and alignments with different approaches. In this example, the time series are translated which makes DTW variants and DTW-GI (rotation) fail, whilst GDTW is invariant to all isometries, and is thus robust to such transformation.

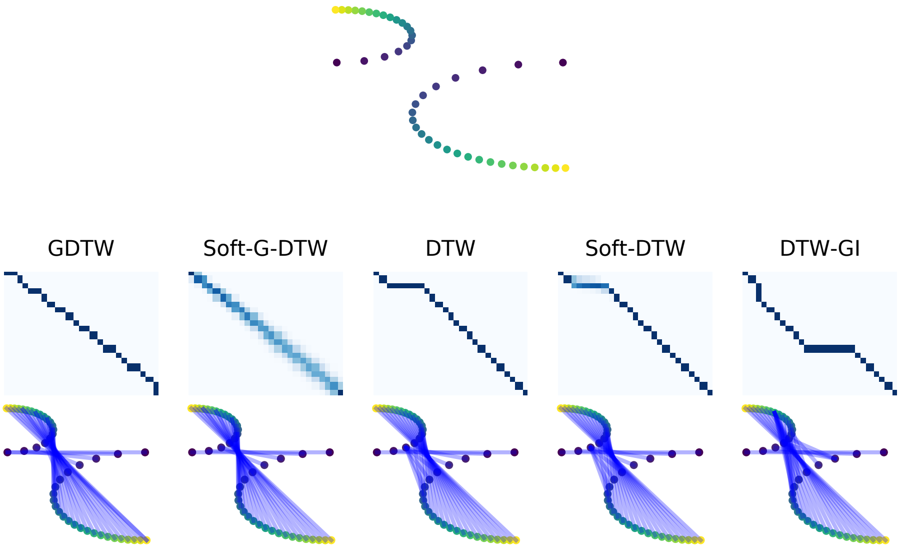


Figure 12: Two time series (top) along with alignment matrices (middle) and alignments with different approaches. In this example, the time series are rotated and translated which makes DTW variants and DTW-GI (rotation) fail, whilst GDTW is invariant to all isometries, and is thus robust to such transformations.