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*

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

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.

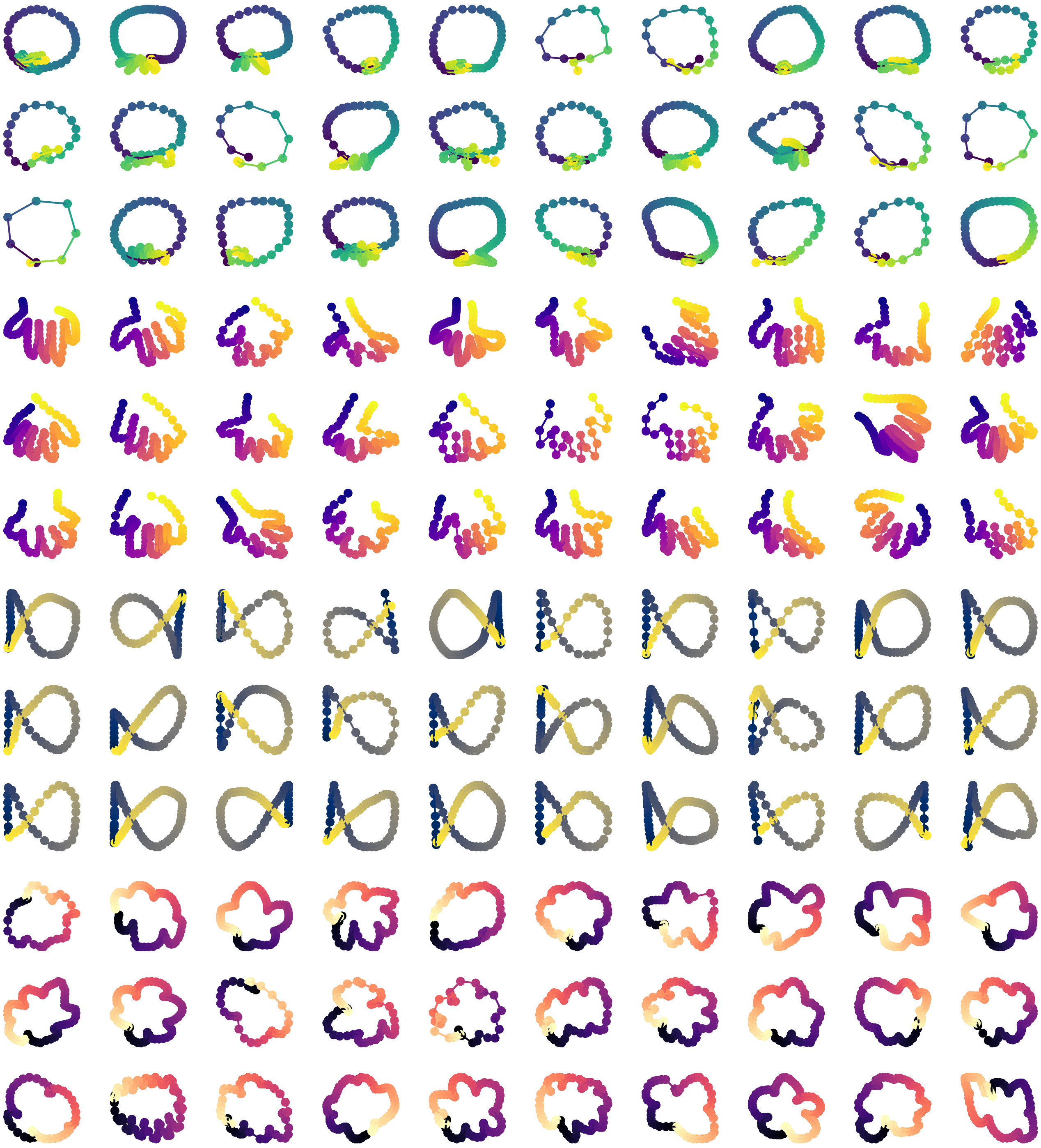

Figure 8: Quickdraw datasets, with classes *blueberries*, *hands*, *fishes*, *clouds*.

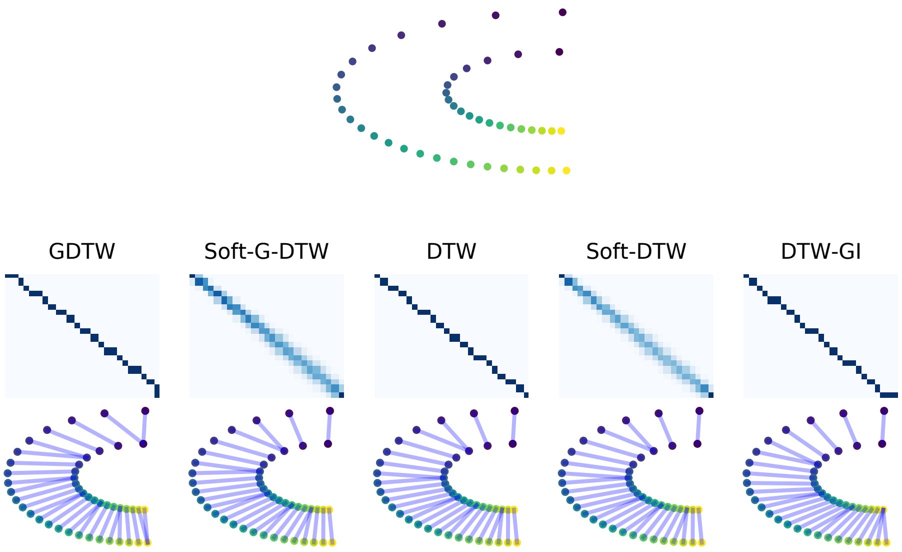

Figure 9: Two time series (top) along with alignment matrices (middle) and alignments with different approaches. In this example, all methods provide a sensible alignment because the time series are on the same axis of rotation and close in the ground space.

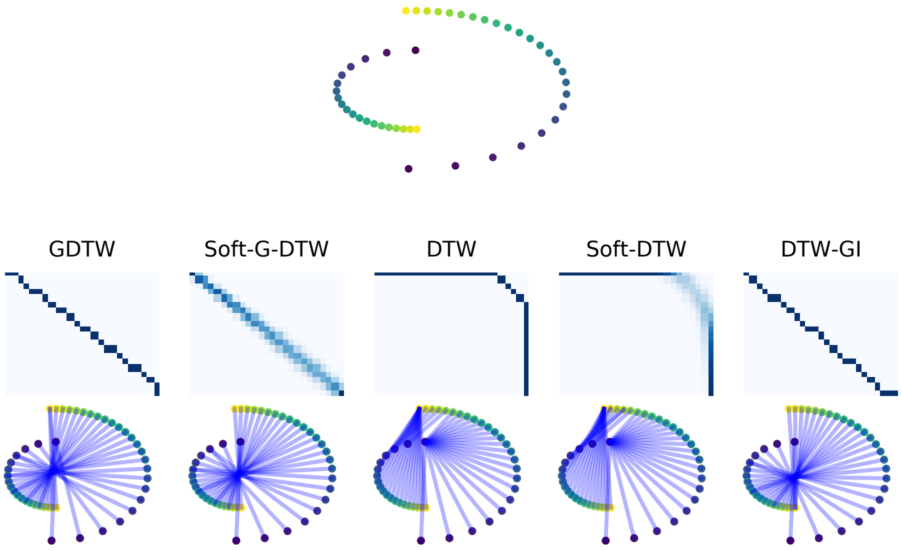

Figure 10: Two time series (top), alignment matrices (middle) and alignments with different approaches. In this example, the time series are not on the same rotation axis which makes DTW variants fail, whilst GDTW and DTW-GI (rotation) provide good alignments due to rotational invariance.

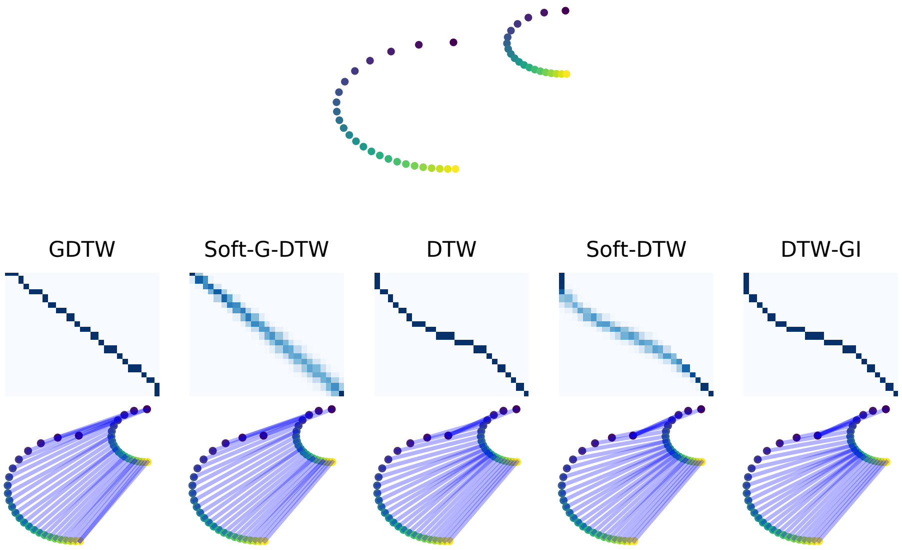

Figure 11: Two time series (top) along with alignment matrices (middle) and alignments with different approaches. In this example, the time series are translated which makes DTW variants and DTW-GI (rotation) fail, whilst GDTW is invariant to all isometries, and is thus robust to such transformation.

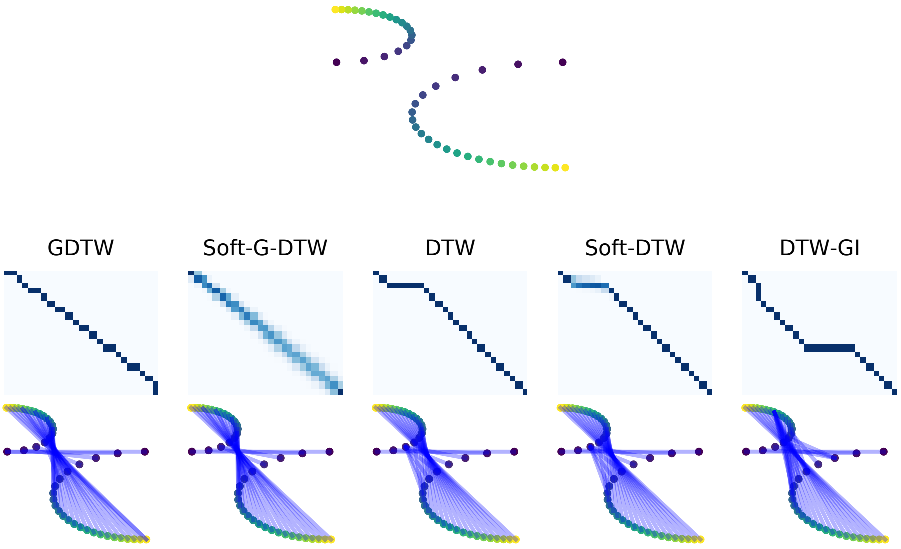

Figure 12: Two time series (top) along with alignment matrices (middle) and alignments with different approaches. In this example, the time series are rotated and translated which makes DTW variants and DTW-GI (rotation) fail, whilst GDTW is invariant to all isometries, and is thus robust to such transformations.