Spatio-Temporal Alignments: Optimal transport through space and time

Hicham Janati
Inria Saclay & ENSAE

Marco Cuturi
Google Brain & ENSAE

Alexandre Gramfort
Inria Saclay

Abstract

Comparing data defined over space and time is notoriously hard. It involves quantifying both spatial and temporal variability while taking into account the chronological structure of the data. Dynamic Time Warping (DTW) computes a minimal cost alignment between time series that preserves the chronological order but is inherently blind to spatio-temporal shifts. In this paper, we propose Spatio-Temporal Alignments (STA), a new differentiable formulation of DTW that captures spatial and temporal variability. Spatial differences between time samples are captured using regularized Optimal transport. While temporal alignment cost exploits a smooth variant of DTW called soft-DTW. We show how smoothing DTW leads to alignment costs that increase quadratically with time shifts. The costs are expressed using an unbalanced Wasserstein distance to cope with observations that are not probabilities. Experiments on handwritten letters and brain imaging data confirm our theoretical findings and illustrate the effectiveness of STA as a dissimilarity for spatio-temporal data.

1 Introduction

To discriminate between two sets of observations, one must find an appropriate metric that emphasizes their differences. The performance of any machine learning model is thus inherently conditioned by the discriminatory power of the metrics it is built upon. Yet, designing the *best* metric for the application at hand is not an easy task. A *good* metric must take into account the structure of its inputs. Here we propose a differentiable metric for spatio-temporal data.

Spatio-temporal data Spatio-temporal data consist of time series where each time sample is multivariate and lives in a certain coordinate system equipped with a natural distance. Such a coordinate system can correspond to 2D or 3D positions in space, pixel positions etc. This setting is encountered in several machine learning problems. Multi-target tracking for example, involves the prediction of the time indexed positions of several objects or particles (Doucet et al., 2002). In brain imaging, magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) yield measurements of neural activity in multiple positions and at multiple time points (Gramfort et al., 2011). Quantifying spatio-temporal variability in brain activity can allow to compare different clinial populations. In traffic dynamics studies, several public datasets report the tracked movements of pedestrians and cars such as the NYC taxi data (Taxi and Commission, 2019).

Optimal transport Recently, optimal transport has gained considerable interest from the machine learning and signal processing community (Peyré and Cuturi, 2018). Indeed, when data are endowed with geometrical properties, Optimal transport metrics (a.k.a the Wasserstein distance) can capture spatial variations between probability distributions. Given a transport cost function – commonly referred to as ground metric – the Wasserstein distance computes the optimal transportation plan between two measures. Its heavy computational cost can be significantly reduced by using entropy regularization (Cuturi, 2013). Besides, when measures are not normalized, it is possible to use the unbalanced optimal transport formulation of Chizat et al. (2017), which allows to compute the entropy regularized Wasserstein distance using Sinkhorn's algorithm with minor modifications. To take into account the temporal dimension, one could define the ground metric as a combination of spatial and temporal shifts similarly to the definition of TL^p distances (Thorpe et al., 2017).

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

This method however ignores the chronological order of the data and requires a tuning parameter to settle the tradeoff between spatial and temporal transport cost. Instead, one can make use of the dynamic time warping (DTW) framework.

Dynamic time warping Given a pairwise distance matrix between all time points of two time series of respective lengths m, n, DTW computes the minimum-cost alignment between the time series (Sakoe and Chiba, 1978) while preserving the chronological order of the data. Indeed, the DTW optimization problem is constrained on alignments where no temporal back steps are allowed. It can be seen as an OT-like problem where the transport plan must not respect the marginal constraints but instead is a binary matrix with at least one non-zero entry per line and per column, and where the cumulated non-zero path is formed by $\rightarrow, \downarrow, \searrow$ steps exclusively. However, the binary nature of this set makes the DTW loss non-differentiable which is a major limitation when DTW is used as a loss function. To circumvent this issue, several authors introduced smoothed versions of DTW (Saigo et al., 2004; Cuturi, 2011; Cuturi and Blondel, 2017). Instead of selecting *the* minimum cost alignment, Global Alignment Kernels (GAK) Saigo et al. (2004); Cuturi (2011) compute a weighted cost on the whole set of possible alignments. Similarly, the soft-minimum generalization approach of Cuturi and Blondel (2017) – called soft-DTW – provides a similar framework to that of GAK where gradients can easily be computed used a backpropagation of Bellman's equation (Bellman, 1952).

Our Contributions Our contributions are twofold. First, we show that, contrarily to DTW that is blind to time shifts, soft-DTW captures temporal shifts with a quadratic lower bound. Second, we propose to use a divergence based on unbalanced optimal transport as a cost for the soft-DTW loss function. The resulting distance-like function is differentiable and can capture both spatial and temporal differences. We call it Spatio-Temporal Alignment (STA). Since the optimal temporal alignment between two time series is computed by minimizing the overall *spatial* transportation cost, this formulation leads to an intuitive metric to compare time series of spatially defined samples while taking into account the chronological structure of the data. We experimentally illustrate the relevance of STA on clustering tasks of brain imaging and handwritten letters datasets.

Structure Section 2 provides some background material on optimal transport and dynamic time warping. We show in section 3 that soft-DTW increases at least quadratically with temporal shifts. In Section 4 we introduce the proposed STA dissimilarity. Finally, Section 5 illustrates the potential applications of STA using several experiments.

Notation We denote by $\mathbb{1}_p$ the vector of ones in $\mathbb{R}^p$ and by $[\![q]\!]$ the set $\{1, \dots, q\}$ for any integer $q \in \mathbb{N}$. The set of vectors in $\mathbb{R}^p$ with non-negative (resp. positive) entries is denoted by $\mathbb{R}^p_+$ (resp. $\mathbb{R}^p_{++}$). On matrices, log, exp and the division operator are applied element-wise. We use $\odot$ for the element-wise multiplication between matrices or vectors. If $\mathbf{X}$ is a matrix, $\mathbf{X}_{i.}$ denotes its i^{th} row and $\mathbf{X}_{.j}$ its j^{th} column. We define the Kullback-Leibler (KL) divergence between two positive vectors by $\text{KL}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \log(\mathbf{x}/\mathbf{y}) \rangle + \langle \mathbf{y} - \mathbf{x}, \mathbb{1}_p \rangle$ with the continuous extensions $0 \log(0/0) = 0$ and $0 \log(0) = 0$. We also make the convention $\mathbf{x} \neq 0 \Rightarrow \text{KL}(\mathbf{x}|0) = +\infty$. The entropy of $\mathbf{x} \in \mathbb{R}^n$ is defined as $H(\mathbf{x}) = -\langle \mathbf{x}, \log(\mathbf{x}) - \mathbb{1}_p \rangle$. The same definition applies for matrices with an element-wise double sum. The feasible set of binary matrices of $\mathbb{R}^{m \times n}$ where only $\rightarrow, \downarrow, \searrow$ movements are allowed is denoted by $\mathcal{A}_{m,n}$.

2 Background on Optimal transport and soft-DTW

2.1 Unbalanced Optimal transport

Entropy regularization Consider a finite metric space (E, d) where $E = \{1, \dots, p\}$. Let $\mathbf{M}$ be the matrix where $\mathbf{M}_{ij}$ corresponds to the distance d between entry i and j. Let $\mathbf{x}, \mathbf{y}$ be two normalized histograms on E ($\mathbf{x}^\top \mathbb{1} = \mathbf{y}^\top \mathbb{1} = 1$). Assuming that transporting a fraction of mass $\mathbf{P}_{ij}$ from i to j is given by $\mathbf{P}_{ij}\mathbf{M}_{ij}$, the total cost of transport is given by $\langle \mathbf{P}, \mathbf{M} \rangle = \sum_{ij} \mathbf{P}_{ij}\mathbf{M}_{ij}$. The Wasserstein distance is defined as the minimum of this total cost with respect to $\mathbf{P}$ on the polytope $\mathcal{P} = \{\mathbf{P} \in \mathbb{R}^{p \times p}_+, \mathbf{P}\mathbb{1} = \mathbf{x}, \mathbf{P}^\top \mathbb{1} = \mathbf{y}\}$ (Kantorovic, 1942). Entropy regularization was introduced by Cuturi (2013) to propose a faster and more robust alternative to the direct resolution of the linear programming problem. Formally, this accounts to minimizing the loss $\langle \mathbf{P}, \mathbf{M} \rangle - \varepsilon H(\mathbf{P})$ where $\varepsilon > 0$ is a regularization hyperparameter. Up to a constant, this problem is equivalent to:

$$\min_{\mathbf{P} \in \mathcal{P}} \varepsilon \text{KL}(\mathbf{P}, e^{-\frac{\mathbf{M}}{\varepsilon}}) \ , \tag{1}$$

which can be solved using Sinkhorn's algorithm.

Unbalanced Wasserstein To cope with unbalanced inputs, Chizat et al. (2017) proposed to relax the marginal constraints of the polytope $\mathcal{P}$ using a Kullback-Leibler divergence. Given a hyperparameter

$\gamma > 0$:

$$W(\mathbf{x},\mathbf{y}) = \min_{\mathbf{P}\in\mathbb{R}_+{}^{p\times p}} \varepsilon\mathrm{KL}(\mathbf{P}|e^{-\frac{\mathbf{M}}{\varepsilon}}) + \gamma\mathrm{KL}(\mathbf{P}\mathbb{1}|\mathbf{x}) + \gamma\mathrm{KL}(\mathbf{P}^\top\mathbb{1}|\mathbf{y}) \ . \tag{2}$$

While the first term minimizes transport cost, the added Kullback-Leibler divergences penalize for mass discrepancies between the transport plan and the input unnormalized histograms. Problem (2) can be solved using the following proposition.

Proposition 1. *Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^p_+$. The unbalanced Wasserstein distance is obtained from the dual problem:*

$$W(\mathbf{x},\mathbf{y}) = \max_{u,v\in\mathbb{R}^p} -\gamma\langle\mathbf{x}, e^{-\frac{u}{\gamma}} - 1\rangle - \gamma\langle\mathbf{y}, e^{-\frac{v}{\gamma}} - 1\rangle - \varepsilon\langle e^{\frac{u\oplus v}{\varepsilon}} - 1, e^{-\frac{M}{\varepsilon}}\rangle \ . \tag{3}$$

Moreover, with the change of variables: $\omega = \frac{\gamma}{\gamma+\varepsilon}$, $\mathbf{K} = e^{-\frac{\mathbf{M}}{\varepsilon}}, \mathbf{a} = e^{\frac{u}{\varepsilon}}, \mathbf{b} = e^{\frac{v}{\varepsilon}}$, the optimal dual points are the solutions of the fixed point problem:

$$\mathbf{a} = \left(\frac{\mathbf{x}}{K\mathbf{b}}\right)^\omega \quad , \quad \mathbf{b} = \left(\frac{\mathbf{y}}{K^\top\mathbf{a}}\right)^\omega \tag{4}$$

and the optimal transport plan is given by:

$$(\mathbf{P}_{ij}) = (\mathbf{a}_i\mathbf{K}_{ij}\mathbf{b}_j). \tag{5}$$

PROOF. Since the conjugate of the linear operator $G : \mathbf{P} \mapsto (\mathbf{P}\mathbb{1}, \mathbf{P}^\top\mathbb{1})$ is given by $G^\star : (u,v) \mapsto u \oplus v$, the Fenchel duality theorem leads to (2). The dual loss function is concave and goes to $-\infty$ when $\|u,v\| \to +\infty$, canceling its gradient yields (57). Finally, since the primal problem is convex, strong duality holds and the primal-dual relationship gives (5). See (Chizat et al., 2017) for a detailed proof. □

Solving the fixed point problem (57) is equivalent to alternate maximization of the dual function (3). Starting from two vectors $\mathbf{a}, \mathbf{b}$ set to $\mathbb{1}$, the algorithm iterates through the scaling operations (57). This is a generalization of the Sinkhorn algorithm which corresponds to $\omega = 1$ or $\gamma = +\infty$.

Corollary 1. *Let $\mathbf{x} \in \mathbb{R}^p_+$. The associated optimal dual scalings $\mathbf{a}$, $\mathbf{b}$ to computing $W(\mathbf{x},\mathbf{x})$ are given by the solution of the fixed point problem: $\mathbf{b} = \mathbf{a} = \left(\frac{\mathbf{y}}{K\mathbf{a}}\right)^\phi$*

PROOF. The symmetry of the dual problem (3) with $\mathbf{x} = \mathbf{y}$ implies immediately that $\mathbf{a} = \mathbf{b}$. Proposition 1 gives the fixed point equation.

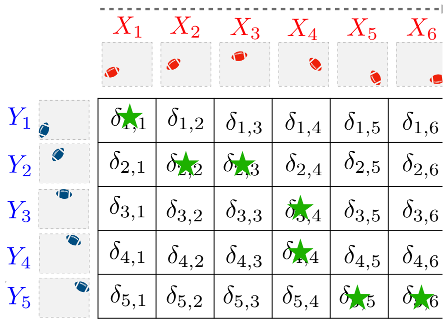
2.2 Soft Dynamic Time Warping

Forward recursion Let $\mathbf{x} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_{T_1}^\top)^\top \in \mathbb{R}^{T_1,p}$ and $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_{T_2}^\top)^\top \in \mathbb{R}^{T_2,p}$ be two time series of respective lengths T_1, T_2 and dimension p. The set of all feasible alignments in a (T_1, T_2) rectangle is denoted by $\mathcal{A}_{T_1,T_2}$. Given a pairwise distance matrix $\Delta(\mathbf{x},\mathbf{y}) \overset{\text{def}}{=} (\delta(\mathbf{x}_i,\mathbf{y}_j))_{ij}$, soft-DTW is defined as:

$$\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y};\Delta) = \mathrm{softmin}_\beta\{\langle\mathbf{A}, \Delta(\mathbf{x},\mathbf{y})\rangle, \mathbf{A} \in \mathcal{A}_{T_1,T_2}\} \ , \tag{6}$$

where the soft-minimum operator of a set $\mathcal{A}$ with parameter β is defined as:

$$\mathrm{softmin}_\beta(\mathcal{A}) = \begin{cases} -\beta\log\left(\sum_{\mathcal{A}} e^{-a/\beta}\right) & \text{if } \beta > 0 \\ \min_{a\in\mathcal{A}} a \quad \text{if } \beta = 0 \end{cases} \tag{7}$$



	X_1	X_2	X_3	X_4	X_5	X_6
Y_1	$\delta_{1,1}$	$\delta_{1,2}$	$\delta_{1,3}$	$\delta_{1,4}$	$\delta_{1,5}$	$\delta_{1,6}$
Y_2	$\delta_{2,1}$	$\delta_{2,2}$	$\delta_{2,3}$	$\delta_{2,4}$	$\delta_{2,5}$	$\delta_{2,6}$
Y_3	$\delta_{3,1}$	$\delta_{3,2}$	$\delta_{3,3}$	$\delta_{3,4}$	$\delta_{3,5}$	$\delta_{3,6}$
Y_4	$\delta_{4,1}$	$\delta_{4,2}$	$\delta_{4,3}$	$\delta_{4,4}$	$\delta_{4,5}$	$\delta_{4,6}$
Y_5	$\delta_{5,1}$	$\delta_{5,2}$	$\delta_{5,3}$	$\delta_{5,4}$	$\delta_{5,5}$	$\delta_{5,6}$

Fig. 1. Example of Dynamic time warping alignment between two time series of images given a pairwise distance matrix.

Algorithm 1 BP recursion to compute $\mathbf{dtw}_\beta$ (Cuturi and Blondel, 2017)

```
Input: data $\mathbf{x}, \mathbf{y}$ soft-min parameter $\beta$ and distance function $\delta$
Output: $\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y}) = r_{T_1,T_2}$
$r_{0,0} = 0; r_{0,j} = r_{i,0} = \infty$ for $i \in [\![T_1]\!]$ , $j \in [\![T_2]\!]$
for $i = 1$ to $T_1$ do
  for $j = 1$ to $T_2$ do
    $r_{i,j} = \delta(\mathbf{x}_i,\mathbf{y}_j)$ softmin$_\beta(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1})$
  end for
end for
```

Figure 1 illustrates two time series of images and their cost matrix Δ. The path from (1, 1) to (5, 6) is an example of a feasible alignement in $\mathcal{A}_{5,6}$. When $\beta = 0$, the soft-minimum is a minimum and $\mathbf{dtw}_\beta$ falls back to the classical DTW metric. Nevertheless, it can still be computed using the dynamic program of Algorithm 1 with a soft-min instead of min operator.

3 Soft-DTW captures time shifts

Temporal shifts Let $\mathbf{x}$ and $\mathbf{y}$ be two time series. When studying the properties of $\mathbf{dtw}_\beta$, the dimensionality of the time series is irrelevant since it is compressed when computing the cost matrix Δ. Thus, to study temporal shifts, we assume in this section that $\mathbf{x}$ and $\mathbf{y}$ are univariate and belong to $\mathbb{R}^T$. To properly define temporal shifts, we introduce a few preliminary notions. We name the first (respectively, last) time index where $\mathbf{x}$ fluctuates the *onset* (respectively, the *offset*) of $\mathbf{x}$ and denote it by $\text{on}(\mathbf{x})$ (respectively, $\text{off}(\mathbf{x})$). The *fluctuation set* of $\mathbf{x}$ is denoted by $\text{fluc}(\mathbf{x})$ and corresponds to all time indices between the onset and the offset. Formally:

$$\text{on}(\mathbf{x}) = \underset{i \in [\![1, T-1]\!]}{\arg\min} \{\mathbf{x}_{i+1} \neq \mathbf{x}_i\} \tag{8}$$

$$\text{off}(\mathbf{x}) = \underset{i \in [\![1, T-1]\!]}{\arg\max} \{\mathbf{x}_{i+1} \neq \mathbf{x}_i\} \tag{9}$$

$$\text{fluc}(\mathbf{x}) = \{i \in [\![1, T]\!], \text{on}(\mathbf{x}) \leq i \leq \text{off}(\mathbf{x})\} \tag{10}$$

For $\mathbf{x}$ and $\mathbf{y}$ to be temporally shifted with respect to each other, their values must agree both within and outside their (different) fluctuation sets.

Definition 1 (Temporal k-shift)**.** *Let $\mathbf{x}$ and $\mathbf{y}$ be two time series in $\mathbb{R}^T$ and $k \in [\![1, T-1]\!]$. We say that $\mathbf{y}$ is temporally k-shifted with respect to $\mathbf{x}$ and write $\mathbf{y} = \mathbf{x}_{+_k}$ if and only if:*

$$\begin{aligned}
&\text{on}(\mathbf{y}) = \text{on}(\mathbf{x}) + k \\
&\text{off}(\mathbf{y}) = \text{off}(\mathbf{x}) + k \\
&i \leq \text{on}(\mathbf{x}), j \leq \text{on}(\mathbf{y}) \Rightarrow \mathbf{x}_i = \mathbf{y}_j \\
&i \geq \text{off}(\mathbf{x}), j \geq \text{off}(\mathbf{y}) \Rightarrow \mathbf{x}_i = \mathbf{y}_j \\
&i \in \text{fluc}(\mathbf{x}), j \in \text{fluc}(\mathbf{y}), |i - j| = k \Rightarrow \mathbf{x}_i = \mathbf{y}_j \ .
\end{aligned} \tag{11}$$

An example of a temporal 50-shift is illustrated in Figure 2. The heatmap of the squared Euclidean cost matrix Δ shows three rectangular white areas where all alignments A, B and C have the same cost of 0. Since $\mathbf{dtw}_0$ is defined as the minimum of all alignment costs, all these paths are equivalent. Temporal k-shifts change the set of alignments with cost 0 but do not change the $\mathbf{dtw}_0$ value. However, when $\beta > 0$, $\mathbf{dtw}_\beta$ computes a weighted sum of all possible paths, which is affected by temporal shifts by including the number of equivalent paths. The cardinality of $\mathcal{A}_{m,n}$ is known as the Delannoy number $D(m-1, n-1)$ (Cuturi, 2011). For the sake of convenience, we consider the shifted Delannoy sequence starting at $n = m = 1$ so that: $\text{card}(\mathcal{A}_{m,n}) = D_{m,n}$. If β is positive but small enough, the alignements with 0 cost dominate the $\mathbf{dtw}_\beta$ logsumexp. This leads to proposition 2.

Definition 2 (Delannoy sequence)**.** *The Delannoy number $D_{m,n}$ corresponds to the number of paths from $(1,1)$ to (m,n) in a $(m \times n)$ lattice where only $\rightarrow, \downarrow, \searrow$ movements are allowed. It can also be defined with the recursion $\forall m, n \in \mathbb{N}^\star$:*

$$D_{1,n} = D_{m,1} = 1 \tag{12}$$

$$D_{m+1,n+1} = D_{m,n+1} + D_{m+1,n} + D_{m,n} \ . \tag{13}$$

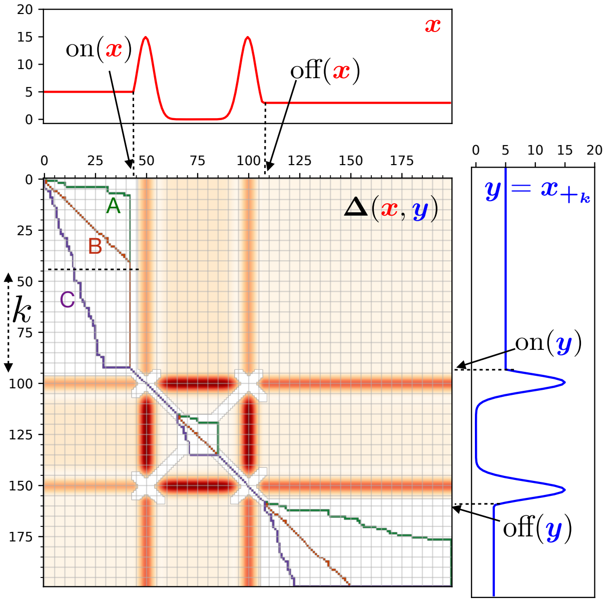


Fig. 2. Example of 3 DTW alignment paths (A, B and C) between $\mathbf{x}$ and $\mathbf{y} = \mathbf{x}_{+_k}$ with a temporal 50-shift. The heatmap of the distance matrix Δ shows (white) rectangles where all paths A, B, C have an equal DTW cost of 0. These areas correspond to time durations where $\mathbf{x}$ and $\mathbf{x}_{+_k}$ are constant. It is noteworthy that when shifting one time series, among the areas crossed by the alignments A, B, C, only the two white rectangles outside the fluctuation set change in size.

Proposition 2. *Let $k \in [\![1, T-1]\!]$, let $m = \text{on}(\mathbf{x})$ and $m' = T - \text{off}(\mathbf{x})$. Let $\mu = \min_{i,j}\{\Delta(\mathbf{x}, \mathbf{x})_{ij} | \Delta(\mathbf{x}, \mathbf{x})_{ij} > 0\}$. If $0 < \beta \leq \frac{\mu}{\log(3TD_{T,T})}$:*

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+_k}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq \beta \log\left(\frac{D_{m,m} D_{m',m'}}{D_{m+k,m} D_{m'-k,m'}}\right) - \frac{\beta}{3T} \tag{14}$$

Sketch of proof. When β is small, the logsumexp in the $\mathbf{dtw}_\beta$ is dominated by the number of alignments with 0 cost. This number is given by: $D_{\text{on}(\mathbf{x}),\text{on}(\mathbf{y})} \Omega D_{T-\text{off}(\mathbf{x}), T-\text{off}(\mathbf{y})}$, where Ω is the number of 0 cost alignments within the cross product of the fluctuation sets. However, temporal shifts do not change Ω but only change the outermost sets. For instance, considering the example of Figure 2 one can see that only rectangles outside the fluctuation set are affected. Therefore, Ω cancels out in the first term of (14). Using the upper bound on β, we derive the second term. The full proof is provided in the supplementary materials. □

Quadratic lower bound The purpose of the rest of the section is to find a lower bound of the right side of

(14) that depends on k. To do so, we incrementally upper bound the off-diagonal Delannoy number $D_{m,m+k}$ with its left and bottom neighbors. When $k = 1$, the following Lemma happens to be crucial to derive the lower bound.

Lemma 1 (Bounded growth)**.** *Let $c = 1 + \sqrt{2}$ and $m \geq 1$. The central (diagonal) Delannoy numbers $D_m = D_{m,m}$ verify:*

$$D_{m+1} \leq c^2 D_m \tag{15}$$

PROOF. The proof is provided in the supplementary materials.

Proposition 3. *Let $c = 1 + \sqrt{2}$. $\forall m, i \in \mathbb{N}^\star$:*

$$D_{m,m+i} \leq c\Phi_{m,i} D_{m,m+i-1} \tag{16}$$

$$c\Psi_{m,i} D_{m,m+i} \leq D_{m+1,m+i} \tag{17}$$

Where

$$\begin{cases} \Phi_{m,i} = 1 - \frac{(1-\frac{1}{c})(i-1)+\frac{1}{c}}{m+i-1} \\ \Psi_{m,i} = 1 + \frac{(1-\frac{1}{c})(i-1)}{m} \end{cases}$$

SKETCH OF PROOF. We prove both statements jointly with a double recurrence reasoning. The initializing for $i = 1$ is immediately obtained using the bounded growth Lemma 1. To show the induction step, we rely on the recursion equation (13). For the sake of brevity, the full proof is provided in the supplementary materials.

By applying proposition 3 to all $i \in [\![1, k]\!]$, the product of all the obtained inequalities leads to a bound on the right side of proposition 2:

Proposition 4. *Let $k \in [\![1, T-1]\!]$, let $m = \mathrm{on}(\mathbf{x})$ and $m' = T - \mathrm{off}(\mathbf{x})$. Using the notations of proposition 3 for Φ and Ψ:*

$$\log\left(\frac{D_{m,m} D_{m',m'}}{D_{m+k,m} D_{m'-k,m'}}\right) \geq \log\left(\prod_{i=1}^{k} \frac{\Psi_{m'-i,i}}{\Phi_{m,i}}\right) \tag{18}$$

PROOF. Iterating the inequalities of proposition 3, we have on one hand with the first inequality:

$$\frac{D_{m,m}}{D_{m,m+k}} \geq \frac{1}{c^k \prod_{i=1}^{k} \Phi_{m,i}} \quad , \tag{19}$$

and on the other hand with the second inequality:

$$\frac{D_{m+k,m+k}}{D_{m,m+k}} \geq c^k \prod_{i=0}^{k-1} \Psi_{m+i,k-i} = c^k \prod_{i=1}^{k} \Psi_{m+k-i,i} \quad .$$

With the change of variable $m' = m + k$ and the symmetry of Delannoy numbers, we have:

$$\frac{D_{m',m'}}{D_{m',m'-k}} \geq c^k \prod_{i=1}^{k} \Psi_{m'-i,i} \quad . \tag{20}$$

Taking the product of (19) and (20) and the result of proposition 2 concludes the proof. □

Finally, we can now state our main theorem.

Theorem 1. *Let $k \in [\![1, T-1]\!]$, let $m = \mathrm{on}(\mathbf{x})$ and $m' = T - \mathrm{off}(\mathbf{x})$. Let $\mu = \min_{i,j}\{\Delta(\mathbf{x},\mathbf{x})_{ij} | \Delta(\mathbf{x},\mathbf{x})_{ij} > 0\}$. If $0 < \beta \leq \frac{\mu}{\log(3TD_{T,T})}$:*

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+_k}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq \beta\alpha k(k-1) + \beta\rho k \tag{21}$$

Where $\alpha = \frac{2-\sqrt{2}}{2}(\frac{1}{m'} + \frac{1}{m+m'}) > 0$ and $\rho = \frac{3\sqrt{2}-4}{3T} > 0$.

PROOF. Let $a = 1 - \frac{1}{c}$. Developping the bound in proposition 4, we get:

$$\log\left(\prod_{i=1}^{k} \frac{\Psi_{m'-i,i}}{\Phi_{m,i}}\right) = \sum_{i=1}^{k} \log(\Psi_{m'-i,i}) - \log(\Phi_{m,i}) \tag{22}$$

Using the inequality $\frac{x}{1+x} \leq \log(1+x) \leq x$ for $x > -1$ on both logarithms we have, on one hand:

$$\begin{aligned} \log(\Psi_{m'-i,i}) &= \log\left(1 + \frac{a(i-1)}{m'-i}\right) \\ &\geq \frac{a(i-1)}{m'-i+a(i-1)} = \frac{a(i-1)}{m' - \frac{i}{c} - a} \\ &\geq \frac{a(i-1)}{m'} \end{aligned} \tag{23}$$

and on the other hand:

$$\begin{aligned} -\log(\Phi_{m,i}) &= -\log\left(1 - \frac{a(i-1)+\frac{1}{c}}{m+i-1}\right) \\ &\geq \frac{a(i-1)+\frac{1}{c}}{m+i-1} \geq \frac{a(i-1)+\frac{1}{c}}{m+m'} \\ &\geq \frac{a(i-1)}{m+m'} + \frac{1}{cT} \quad . \end{aligned} \tag{24}$$

Finally, combining equations (23) and (24), the formula $\sum_{i=1}^{k}(i-1) = \frac{k(k-1)}{2}$ and adding the term $-\frac{\beta}{3T}$ of (14) leads the desired quadratic function. □

We illustrate these bounds experimentally with the example of Figure 2 with $T = 400$ to allow for larger temporal shifts and $\beta = 0.1$. Figure 3 shows that $\mathbf{dtw}_\beta$ is indeed polynomial in k; the quadratic bound is sufficient as an approximation for the result of theorem 1. Experimentally, we notice that the assumption on β is too restrictive. Indeed, the comparison empirically holds for larger values of β which may be desirable in practice to capture more temporal differences. The corresponding figures are provided in the appendix.

4 Spatio-Temporal Alignments

Unbalanced Sinkhorn divergence To capture spatial variability, we propose to use a cost function

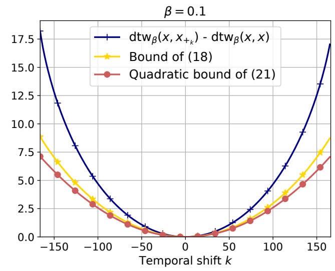


Fig. 3. Illustration of the bounds of proposition 4 and theorem 1 with $\beta = 0.1$ and $T = 400$. The time series $\mathbf{x}$ is a centered version of the one displayed in Figure 2.

based on the unbalanced Wasserstein distance W. Since $W(\mathbf{x}, \mathbf{x}) \neq 0$, the resulting metric would fail to identify identical samples. Similarly to the introduction of Sinkhorn divergences for the balanced case (Genevay et al., 2018), we define the unbalanced Sinkhorn divergence S between two histograms in $\mathbb{R}^p_+$ as:

$$S(\mathbf{x}, \mathbf{y}) = W(\mathbf{x}, \mathbf{y}) - \frac{1}{2}\left(W(\mathbf{x}, \mathbf{x}) + W(\mathbf{y}, \mathbf{y})\right) \tag{25}$$

The proposed dissimilarity – *Spatio-Temporal Alignement* – corresponds to the soft-DTW loss with the divergence S as an alignement cost:

Definition 3 (STA)**.** *We define the STA loss as:*

$$\mathbf{sta}_\beta(\mathbf{x}, \mathbf{y}) = \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}; S) \tag{26}$$

Aside from $S(\mathbf{x},\ \mathbf{x}) =\ 0$, we do not know much about S. The rest of this section aims at showing some of its useful properties when $\mathbf{K}$ is positive semi-definite: non-negativity and coercivity. The curvature of S is however harder to analyze. Nonetheless, it is minimized at $\mathbf{x} = \mathbf{y}$ and if $\mathbf{K}$ is, the only stationary points are $\mathbf{x} = \mathbf{y}$.

Non-negativity We show that S is non-negative, we assume that the kernel $\mathbf{K} = e^{-\frac{\mathbf{M}}{\varepsilon}}$ is positive semi-definite. This is the case for example with $\mathbf{M}_{ij} = \|m_i - m_j\|^l$ with $0 < l \leq 2$ if the support of the measures is given by $\{m_1, \dots, m_p \in \mathbb{R}\}$.

Proposition 5. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p_+$. If $\mathbf{K} = e^{-\frac{\mathbf{M}}{\varepsilon}}$ is positive semi-definite:*

$$S(\mathbf{x}, \mathbf{y}) \geq 0$$

Moreover, if $\mathbf{K}$ is positive definite, $S(\mathbf{x}, \mathbf{y}) = 0 \Rightarrow \mathbf{x} = \mathbf{y}$.

PROOF. Let $\mathbf{c}$ and $\mathbf{d}$ denote the solutions of the fixed point problems: $\mathbf{c} = \left(\frac{\mathbf{x}}{K\mathbf{c}}\right)^\omega$ and $\mathbf{d} = \left(\frac{\mathbf{y}}{K\mathbf{d}}\right)^\omega$. With the change of variable $\mathbf{a} = e^{\frac{\mathbf{u}}{\varepsilon}}$ and $\mathbf{b} = e^{\frac{\mathbf{v}}{\varepsilon}}$, let $(\mathbf{a}, \mathbf{b}) \to \mathcal{D}(\mathbf{a}, \mathbf{b})$ denote the dual function of (3). On one hand, by Corollary 1, $W(\mathbf{x}, \mathbf{x}) = \max_{\mathbf{a},\mathbf{b} \in \mathbb{R}^p_+} \mathcal{D}(\mathbf{a}, \mathbf{b}) = \mathcal{D}(\mathbf{c}, \mathbf{c})$. Similarly, $W(\mathbf{y}, \mathbf{y}) = D(\mathbf{d}, \mathbf{d})$. On the other hand, by definition of the max $W(\mathbf{x}, \mathbf{y}) \geq \mathcal{D}(\mathbf{c}, \mathbf{d})$. Therefore:

$$\begin{aligned}
S(\mathbf{x}, \mathbf{y}) &\geq D(\mathbf{c}, \mathbf{d}) - \frac{1}{2}(D(\mathbf{c}, \mathbf{c}) + D(\mathbf{d}, \mathbf{d})) \\
&= \varepsilon\left[-\langle \mathbf{c} \otimes \mathbf{d}, K\rangle + \frac{1}{2}\langle \mathbf{c} \otimes \mathbf{c}, K\rangle + \frac{1}{2}\langle \mathbf{d} \otimes \mathbf{d}, K\rangle\right] \\
&= \varepsilon\left[-\langle \mathbf{c}, K\mathbf{d}\rangle + \frac{1}{2}\langle \mathbf{c}, K\mathbf{c}\rangle + \frac{1}{2}\langle \mathbf{d}, K\mathbf{d}\rangle\right] \\
&= \frac{\varepsilon}{2}\langle \mathbf{c} - \mathbf{d}, \mathbf{K}(\mathbf{c} - \mathbf{d})\rangle \geq 0
\end{aligned}$$

Where the last inequality follows from the positivity of $\mathbf{K}$. If $\mathbf{K}$ is positive definite, the last inequality is strict unless $\mathbf{c} = \mathbf{d}$, in which case the fixed point equations lead to $\mathbf{x} = \mathbf{y}$. □

Coercivity Regardless of the nature of $\mathbf{K}$, we will now show that $S(., \mathbf{y})$ is coercive for any fixed $\mathbf{y}$. To do so, we first show that S only depends on the sums of transported mass:

Proposition 6. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p_+$ and $\mathbf{P}_{\mathbf{x},\mathbf{y}} \in \mathbb{R}^{p \times p}_+$ their associated transport plan, solution of (2). Then:*

$$S(\mathbf{x}, \mathbf{y}) = (\varepsilon + 2\gamma)(\frac{1}{2}\|P_{\mathbf{x},\mathbf{x}}\|_1 + \frac{1}{2}\|P_{\mathbf{y},\mathbf{y}}\|_1 - \|P_{\mathbf{x},\mathbf{y}}\|_1) \tag{27}$$

SKETCH OF PROOF. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p_+$. And let $\mathbf{a}, \mathbf{b}$ the dual scalings associated with the dual problem of $W(\mathbf{x}, \mathbf{y})$. The corresponding primal solution is given by $\mathbf{P}_{ij} = \mathbf{a}_i\mathbf{K}_{ij}\mathbf{b}_j$. Therefore, using the fixed point equations (57), we have: $\|\mathbf{P}_{\mathbf{x},\mathbf{y}}\|_1 = \langle \mathbf{a}, \mathbf{K}\mathbf{b}\rangle = \langle \mathbf{x}, \mathbf{a}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{b}, \mathbf{K}^\top\mathbf{a}\rangle = \langle \mathbf{y}, \mathbf{b}^{-\frac{\varepsilon}{\gamma}}\rangle$. Therefore, at optimality, the dual function (3) is equal to:

$$W(\mathbf{x}, \mathbf{y}) = -(\varepsilon + 2\gamma)\|\mathbf{P}_{\mathbf{x},\mathbf{y}}\|_1 + \gamma(\|\mathbf{x}\|_1 + \|\mathbf{y}\|_1) + \varepsilon\|\mathbf{K}\|_1,$$

Writing $W(\mathbf{x}, \mathbf{x})$ and $W(\mathbf{y}, \mathbf{y})$ in the same way ends the proof. □

To prove that S is coercive, we bound $\|\mathbf{P}_{\mathbf{x},\mathbf{y}}\|_1$ with the ℓ_1 norms of $\mathbf{x}$ and $\mathbf{y}$:

Lemma 2. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p_+$ and $\mathbf{P}_{\mathbf{x},\mathbf{y}} \in \mathbb{R}^{p \times p}_+$ their associated transport plan, solution of (2). Let $\kappa = \min_{i,j} e^{-\frac{M_{ij}}{\gamma}}$. We have the following bounds on the total transported mass:*

$$\kappa\|\mathbf{x}\|_1\|\mathbf{y}\|_1 \leq \|\mathbf{P}_{\mathbf{x},\mathbf{y}}\|_1^{2+\frac{\varepsilon}{\gamma}} \leq p^{\frac{3}{2}}\|\mathbf{x}\|_1\|\mathbf{y}\|_1 \tag{28}$$

SKETCH OF PROOF. Writing the first order optimality condition of (2) links the optimal transport plan $\mathbf{P}$ with the inputs $\mathbf{x}, \mathbf{y}$. The bounds can be easily derived using basic inequalities. For the sake of brevity, the full proof is provided in the supplementary materials.

Proposition 7. *For* $\mathbf{y} \in \mathbb{R}_+^p$*, the function* $\mathbf{x} \mapsto S(\mathbf{x}, \mathbf{y})$ *is coercive.*

PROOF. Lemma 2 and proposition 6, we get, with $\zeta = \frac{1}{2+\frac{\varepsilon}{\gamma}}$:

$$S(\mathbf{x}, \mathbf{y}) \geq \kappa(\|\mathbf{x}\|_1^{2\zeta} + \|\mathbf{y}\|_1^{2\zeta}) - p^{\frac{3}{2}}\|\mathbf{x}\|_1^{\zeta}\|\mathbf{y}\|_1^{\zeta} \tag{29}$$

Therefore: $\|\mathbf{x}\|_1 \to +\infty \Rightarrow S(\mathbf{x}, \mathbf{y}) \ \to \ +\infty$ □

Differentiability $W(., \mathbf{y})$ is differentiable, and its gradient is given by $\gamma(1 - \mathbf{a}^{-\frac{\varepsilon}{\gamma}})$ where $\mathbf{a}$ is the solution of the fixed equation (57) (Feydy et al., 2017). Thus, S is also differentiable. If $\mathbf{K}$ is positive semi-definite then $S \geq 0$ and thus, from the following proposition we conclude that all its stationary points are minimizers:

Proposition 8. *Let* $\mathbf{y}, \mathbf{x} \in \mathbb{R}_{++}^{n \times p}$ *be a stationary point of* S *i.e* $\nabla S(\mathbf{x}, \mathbf{y}) = (\mathbf{0}, \mathbf{0})$*. Then,* $S(\mathbf{x}, \mathbf{y}) = 0$*. Moreover, if* $\mathbf{K}$ *is positive definite, then* $\mathbf{x} = \mathbf{y}$*.*

PROOF. the proof is provided in the appendix.

Optimal transport hyperparameters The unbalanced W metric is defined by two hyperparameters: ε and γ. On one hand, higher values of ε increase the curvature of the minimized loss function thereby accelerating the convergence of Sinkhorn's algorithm. This gain in speed is however at the expense of entropy blurring of the transport plan. On the other hand, $\varepsilon \to 0$ leads to a well documented numerical instability that can be mitigated using log-domain stabilization (Peyré and Cuturi, 2018). Here we set ε to the lowest stable value. A practical scale is provided by taking values of ε proportional to $\frac{m}{p}$ where m is the median of the ground metric $\mathbf{M}$. The marginals parameter γ must be large enough to guarantee transportation of mass. When $\gamma \to 0$, the optimal transport plan $\mathbf{P}^\star \to \mathbf{K}$. Large γ however slows down the convergence of Sinkhorn's algorithm, especially if the input histograms have significantly different total masses. We set γ at the largest value guaranteeing a minimal transport mass using the heuristic proposed in (Janati et al., 2019).

Complexity analysis As shown by Algorithm 1, soft-DTW is quadratic in time. Computing the Sinkhorn divergence matrix is quadratic in p. Moreover, when the time series are defined on regular grids such as images, one could benefit from spatial Kernel separability as introduced in (Solomon et al., 2015). This trick allows to reduce the complexity of Sinkhorn on 2D data from $O(p^2)$ to $O(p^{\frac{3}{2}})$. Moreover, to leverage fast matrix products on GPUs, computing each of the matrices $(W(\mathbf{x}_i, \mathbf{y}_j))_{ij}$ (resp. $(W(\mathbf{x}_i, \mathbf{x}_i))_i, (W(\mathbf{y}_j, \mathbf{y}_j))_j$) can be done in a $(n \times m)$ parallel version of Sinkhorn's algorithm, where each kernel convolution $\mathbf{K}\mathbf{v}$, $\mathbf{K}^\top\mathbf{u}$ is applied to all nm (resp. m, n) dual variables at once.

Signed data The divergence S is defined for non-negative signals only which can be encountered in practice as non-normalized intensities. Yet, one can easily extend $\mathbf{sta}_\beta$ to signed data by talking the absolute values of the signals. Or computing S on positive and negative parts separately before averaging.

5 Experiments

Our main theoretical result states that $\mathbf{dtw}_\beta$ captures temporal shifts only if $\beta > 0$. Moreover, with the unbalanced Wasserstein divergence as a cost, $\mathbf{sta}_\beta$ should capture both spatial and temporal variability. We illustrate this in a brain imaging simulation and a clustering problem of handritten letters.

5.1 Brain imaging

Brain imaging data recordings report the brain activity both in space and time. Thanks to their high temporal resolution, Electroencephalography and Magnetoencephalography can capture response latencies in the order of a millisecond. Abnormal differences in latency, amplitude and topography of brain signals are important biomarkers of several conditions of the central nervous system such as multiple sclerosis (Whelan et al., 2010) or amblyopia (Sokol, 1983). We argue here that $\mathbf{sta}_\beta$ can aggregate all these differences in a meaningful dissimilarity score. To illustrate this, we use the average brain surface derived from real MRI scans and provided by the FreeSurfer software. We compute a triangulated mesh of 642 vertices on the left hemisphere and simulate 4 types of signals as follows. We set $T = 20$ and select 2 activation time points $t_1 = 5$ and $t_2 = 15$. We also select two brain regions in the visual cortex given by FreeSurfer's segmentation known as *V1* (primary visual cortex) and *MT* (middle temporal visual area) which are defined on 17 and 8 vertices respectively. Each generated time series peaks at t_1 or t_2, in a random vertex in V1 or MT with a random amplitude between 1 and 3. For the signals to be more realistic, we apply a Gaussian filter along the temporal and the spatial axes. Examples of the generated data are displayed in Figure 4. We generate $N = 200$ samples (50 per time point / brain region) and compute the pairwise dissimilarity matrices $\mathbf{dtw}_\beta$ and $\mathbf{sta}_\beta$ with $\beta = 0$ and $\beta = 0.1$. Figure 5 shows the t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008; Pedregosa et al., 2011) of the data. As expected, $\mathbf{dtw}_\beta$ cannot capture spatial variability regardless of β. With $\beta = 0$, $\mathbf{sta}_\beta$ separates the data according to the brain region only. Only with positive β can $\mathbf{sta}_\beta$ identify all four groups. Computing the full $\mathbf{sta}_\beta$ dissimilarity matrix performed $\frac{1}{2}N(N+1) \times T^2 = 8040000$

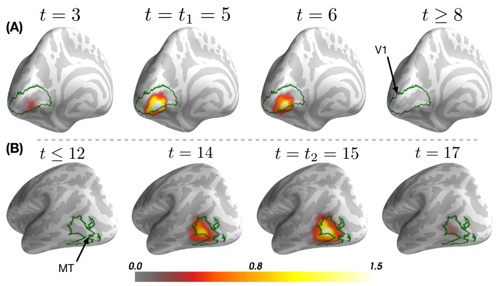


Fig. 4. Two examples of the simulated time series. **(A)** brain signal in V1 with a peak at $t = t_1$. **(B)** brain signal in MT with a peak at $t = t_2$. The borders of the brain regions V1 and MT are highlighted in green.

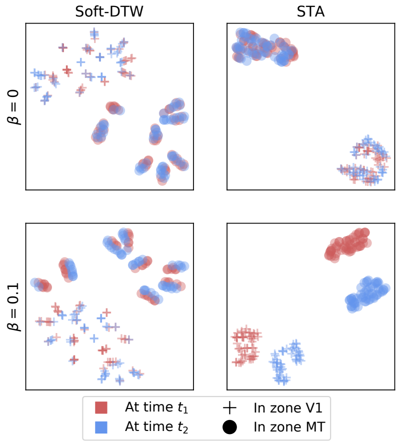


Fig. 5. t-SNE visualization of the simulated brain signals in two different regions, at two different time instants. With $\beta > 0$, $\mathbf{sta}_\beta$ can discriminate between all four groups.

Sinkhorn loops between 642 dimensional inputs. The whole experiment completed in 10 minutes on our DGX-1 station. Python code and data can be found in https://github.com/hichamjanati/spatio-temporal-alignements.

5.2 Handwritten letters

To evaluate the discriminatory power of STA with real data, we use a publicly available dataset of handwritten letters where the position of a pen are tracked in time (Williams et al., 2006). We subsample the data both spatially and temporally so as to keep 10 time points of (64×64) images for each time series. Each image can thus be seen as a screenshot at a certain time during the writing motion. Figure 6 shows an example of two data samples of the letter "g". We consider clustering 140 samples of 7 different letters – ‘a‘ to ‘h‘ – (20 samples per letter; ’f’ was not collected in the data) with a t-SNE embedding using STA as a dissimilarity function. To speed up computation, we compute all pairwise δ distances between all images of all samples on multi-platform GPUs. Carrying out STA afterwards amounts to finding optimal assignments independently for each pair of time series. We compare STA with soft-DTW with the same $\beta = 0.1$. Figure 7 shows that the choice of the dissimilarity is crucial: the spatial variability captured by the Wasserstein divergence is key to accurately discriminate between the samples. In this experiment, we noticed that the choice of β almost did not affect the results. Given that all letters were written by the same person, all motions have similar speeds. Results with various values of β are displayed in the supplementary materials.

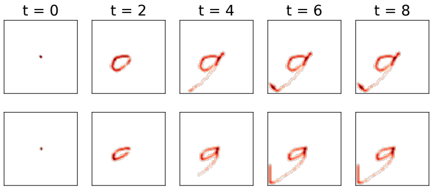


Fig. 6. Examples of time series in the handwritten letters dataset corresponding to the letter "g". At each time point, $\mathbf{x}_i$ corresponds to an image of the current state of the drawing.

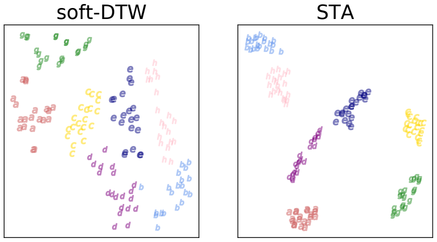


Fig. 7. tSNE embeddings of the data. STA (proposed) captures spatial variability.

6 Conclusion

Spatio-temporal data can differ in amplitude and in spatio-temporal structure. Our contributions are twofold. First, we showed that regularized Dynamic time warping is sensitive to temporal variability. Second, we proposed to combine an unbalanced Optimal transport divergence with soft dynamic time warping to define a dissimilarity for spatio-temporal data. The performance of our experiments on simulations and real data confirm our findings and show that our method can identify meaningful spatio-temporal clusters.

Acknowledgements

MC and HJ acknowledge the support of a chaire d’excellence de l’IDEX Paris Saclay. AG and HJ was supported by the European Research Council Starting Grant SLAB ERC-YStG-676943.

References

Bellman, R. (1952). On the theory of dynamic programming. In *Proceedings of the National Academy of Sciences*, volume 38, page 716–719.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2017). Scaling Algorithms for Unbalanced Transport Problems. *arXiv:1607.05816 [math.OC]*.

Cuturi, M. (2011). Fast global alignment kernels. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 929–936, USA. Omnipress.

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Neural Information Processing Systems*.

Cuturi, M. and Blondel, M. (2017). Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*.

Doucet, A., Vo, B. ., Andrieu, C., and Davy, M. (2002). Particle filtering for multi-target tracking and sensor management. In *Proceedings of the Fifth International Conference on Information Fusion.*, volume 1, pages 474–481 vol.1.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. (2017). Optimal transport for diffeomorphic registration. pages 291–299.

Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR.

Gramfort, A., Papadopoulo, T., Baillet, S., and Clerc, M. (2011). Tracking cortical activity from M/EEG using graph cuts with spatiotemporal constraints. *NeuroImage*, 54(3):1930 – 1941.

Janati, H., Cuturi, M., and Gramfort, A. (2019). Wasserstein regularization for sparse multi-task regression. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*. PMLR.

Kantorovic, L. (1942). On the translocation of masses. *C.R. Acad. Sci. URSS*.

Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peyré, G. and Cuturi, M. (2018). Computational Optimal Transport. *arXiv e-prints*.

Saigo, H., Jean-Philippe, Vert, Ueda, N., and Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Sokol, S. (1983). Abnormal evoked potential latencies in amblyopia. *The British journal of ophthalmology*, 67(5):310–314.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11.

Stanley, R. P. (2011). *Enumerative Combinatorics: Volume 1*. Cambridge University Press, New York, NY, USA, 2nd edition.

Taxi, N. Y. N. Y. . and Commission, L. (2019). New york city taxi trip data, 2009-2018.

Thorpe, M., Park, S., Kolouri, S., Rohde, G. K., and Slepčev, D. (2017). A transportation l(p) distance for signal analysis. *Journal of mathematical imaging and vision*, 59(2):187–210.

Whelan, R., Lonergan, R., Kiiski, H., Nolan, H., Kinsella, K., Bramham, J., O'Brien, M., Reilly, R., Hutchinson, M., and Tubridy, N. (2010). A high-density erp study reveals latency, amplitude, and topographical differences in multiple sclerosis patients versus controls. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 121:1420–6.

Williams, B., M.Toussaint, and Storkey., A. (2006). Extracting motion primitives from natural handwriting data. In *ICANN*, volume 2, page 634–643.

A Proofs

A.1 Proof of proposition 2

Let $\mathbf{x} \in \mathbb{R}^T$ be a univariate time series. Using the definitions of section 3, proposition 2 reads:

Proposition A.1. *Let* $k \in [\![1, T-1]\!]$, *let* $m = \mathrm{on}(\mathbf{x})$ *and* $m' = T - \mathrm{off}(\mathbf{x})$. *Let* $\mu = \min_{i,j}\{\Delta(\mathbf{x},\ \mathbf{x})_{ij}\ |\Delta(\mathbf{x},\ \mathbf{x})_{ij} > 0\}$. *If* $0 < \beta \leq \frac{\mu}{\log(3TD_{T,T})}$ *:*

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+_k}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq \beta \log\left(\frac{D_{m,m}D_{m',m'}}{D_{m+k,m}D_{m'-k,m'}}\right) - \frac{\beta}{3T} \tag{30}$$

PROOF. Let us remind that given a pairwise distance matrix $\Delta(\mathbf{x},\mathbf{y})$, the soft-DTW dissimilarity is defined as $\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y}) = -\beta\log\left(\sum_{A\in\mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle}{\beta}}\right)$. The set of all possible costs can be written: $C = \{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle, A \in \mathcal{A}_{T,T}\}$. Dropping duplicates, let $d_0 < d_1, \dots, < d_G$ denote all unique values in C. And finally let n_i be the number of alignments A such that $\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle = d_i$. We have:

$$\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y}) = -\beta\log\left(\sum_{A\in\mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle}{\beta}}\right) = -\beta\log\left(\sum_{i=0}^{G} n_i e^{-\frac{d_i}{\beta}}\right) \ . \tag{31}$$

When $\mathbf{y} = \mathbf{x}$, we have $d_0 = 0$. Isolating the first element of the sum we get:

$$\mathbf{dtw}_\beta(\mathbf{x},\mathbf{x}) = -\beta\log(n_0) - \beta\log\left(1 + \sum_{i=1}^{G}\frac{n_i}{n_0}e^{-\frac{d_i}{\beta}}\right) \leq -\beta\log(n_0) \ . \tag{32}$$

Similarly, when $\mathbf{y}$ is temporally k-shifted with respect to $\mathbf{x}$, we also have $d_0 = 0$. Adding an exponent $'$ on terms that depend on the time series $\mathbf{x}_{+k}$, we have:

$$\begin{aligned}\mathbf{dtw}_\beta(\mathbf{x},\mathbf{x}_{+k}) = -\beta\log(n'_0) - \beta\log\left(1 + \sum_{i=1}^{G}\frac{n'_i}{n'_0}e^{-\frac{d'_i}{\beta}}\right) &\geq -\beta\log(n'_0) - \beta\sum_{i=1}^{G}\frac{n'_i}{n'_0}e^{-\frac{d'_i}{\beta}} \\ &\geq -\beta\log(n'_0) - \beta D_{T,T}e^{-\frac{d'_1}{\beta}}\end{aligned} \tag{33}$$

However, since the set of values taken by $\Delta(\mathbf{x},\mathbf{x})$ and $\Delta(\mathbf{x},\mathbf{x}_{+k})$ are the same, we have $d_i = d'_i$ (but $n_i \neq n'_i$ apriori) and the assumption on β provides:

$$\begin{aligned}&\beta \leq \frac{\mu}{\log(3TD_{T,T})} \\ &\Rightarrow \beta \leq \frac{d_1}{\log(3TD_{T,T})} \\ &\Rightarrow e^{\frac{-d'_1}{\beta}} \leq \frac{1}{3TD_{T,T}} \\ &\Rightarrow -\beta D_{T,T}e^{\frac{-d'_1}{\beta}} \geq -\frac{\beta}{3T}\end{aligned} \tag{34}$$

Combining (32), (33) and (34) leads to:

$$\mathbf{dtw}_\beta(\mathbf{x},\mathbf{x}_{+k}) - \mathbf{dtw}_\beta(\mathbf{x},\mathbf{x}) \geq \beta\log\left(\frac{n_0}{n'_0}\right) - \frac{\beta}{3T} \tag{35}$$

Now let's develop the term $\frac{n_0}{n'_0}$. n'_0 corresponds to the number of equivalent alignments with 0 cost which can be given by $D_{\mathrm{on}(\mathbf{x}),\mathrm{on}(\mathbf{y})}\Omega D_{T-\mathrm{off}(\mathbf{x}),T-\mathrm{off}(\mathbf{y})}$, where Ω is the number of 0 cost alignments within the cross product of the fluctuation sets. However, temporal shifts do not change Ω but only change the outermost sets. For instance, considering the example of Figure 2 one can see that only rectangles outside the fluctuation set are affected. Therefore, Ω cancels out in $\frac{n_0}{n_0'}$ and we get the desired bound. □

A.2 Proof of proposition 3

For the sake of completeness, we start this section by reminding some key results on Delannoy numbers.

Delannoy numbers

We re-define the Delannoy sequence starting from $m = n = 1$ so as to correspond to the number of chronological alignments in the $(1,\,1) \to (m, n)$ lattice: $\mathrm{card}(\mathcal{A}_{m,n}) = D_{m,n}$.

Definition A.1 (Delannoy sequence)**.** *The Delannoy number $D_{m,n}$ corresponds to the number of paths from $(1, 1)$ to (m, n) in a $(m \times n)$ lattice where only $\rightarrow, \downarrow, \searrow$ movements are allowed. It can also be defined with the recursion $\forall m, n \in \mathbb{N}^\star$:*

$$D_{1,n} = D_{m,1} = 1 \tag{36}$$
$$D_{m+1,n+1} = D_{m,n+1} + D_{m+1,n} + D_{m,n} \quad . \tag{37}$$

The central (or diagonal) Delannoy numbers $D_m = D_{m,m}$ verifiy an intersting 2-stages recursion equation:

Proposition A.2 (Stanley (2011))**.** *For $m \geq 2$:*

$$mD_{m+1} = (6m - 3)D_m - (m - 1)D_{m-1} \tag{38}$$

Sketch of proof. The proof of Stanley (2011) is based on the closed form expression of Delannoy numbers: $D_m = \sum_{k=1}^{m} \binom{m,k}{m+k,k}$ and the generating function $\sum_{m=1}^{\infty} D_m x^n = \frac{1}{\sqrt{1-6x+x^2}}$. Taking the derivative of the power series yields the desired recursion equation.

Lemma A.1 (Bounded growth - Lemma 1)**.** *Let $c = 1 + \sqrt{2}$ and $m \geq 2$. The sequence of central Delannoy numbers $D_m = D_{m,m}$ verifies:*

$$P(m) : D_{m+1} \leq c^2 D_m \tag{39}$$

proof. Proof by induction. For $m = 1$, we have $D_2 = 3 \leq (3 + 2\sqrt{2}) = c^2 = c^2 D_1$. Let $m \geq 2$ and assume $P(m)$ is true. From (38) and $P(m)$ we have:

$$\begin{aligned}(m+1)D_{m+2} &= (6m+3)D_{m+1} - mD_m && (40)\\ &\leq (6m + 3 - \frac{m}{c^2})D_{m+1} && (41)\\ &\leq (6 - \frac{1}{c^2})mD_{m+1} && (42)\\ &\leq (6 - \frac{1}{c^2})(m+1)D_{m+1} && (43)\end{aligned}$$

And we also have $1/c^2 = \frac{1}{3+2\sqrt{2}} = 3 - 2\sqrt{2}$, hence $6 - \frac{1}{c^2} = c^2$; we have $P(m + 1)$. □

Proof of proposition 3

Proposition 3 is our most technical contribution, its demonstration requires considerable care. Similarly to bounded growth Lemma A.1, we would like to bound the off-diagonal Delannoy numbers with their closest diagonal numbers with a bound depending on k. We do so incrementally by comparing the off-diagonal number $D_{m,m+k}$ with $D_{m,m+k-1}$ and $D_{m+1,m+k}$. The proposition states:

Proposition A.3 (Proposition 3)**.** *Let $c = 1 + \sqrt{2}$. $\forall m, k \in \mathbb{N}^\star$:*

$$A(m,k): \quad D_{m,m+k} \leq c\Phi_{m,k} D_{m,m+k-1} \tag{44}$$
$$B(m,k): \quad c\Psi_{m,k} D_{m,m+k} \leq D_{m+1,m+k} \tag{45}$$

Where

$$\begin{cases} \Phi_{m,k} = 1 - \frac{(1-\frac{1}{c})(k-1)+\frac{1}{c}}{m+k-1} \\ \Psi_{m,k} = 1 + \frac{(1-\frac{1}{c})(k-1)}{m} \end{cases}$$

It is noteworthy that – since $1 - 1/c = 2 - \sqrt{2} > 0$ – we have for all m, k $\Phi_{m,k} \leq 1$ and $\Psi_{m,k} \geq 1$. When both Ψ and Φ are constant and equal to 1, we get two constant bounds equal to c. The role of Φ and Ψ is to have tighter bounds when k increases. The demonstration is based on an induction reasoning on m. That is, we would like to show for all m the statement: $P(m) : (\forall k \geq 1)\ A(m,k)$ and $B(m,k)$. To assist the reader, we visualize the proof on Figure 8 which describes all the steps of the induction. For the sake of clarity, we isolate the following technical Lemma before proving the proposition.

Lemma A.2. *Let $c = 1 + \sqrt{2}$ and $m, k \geq 1$. The sequences Φ and Ψ verify the inequalities:*

$$c\Psi_{m,k+1}\Phi_{m,k+1} \leq \left(\frac{1}{c} + \Psi_{m,k} + \Phi_{m,k+1}\right) \leq c\Phi_{m+1,k}\Psi_{m,k} \tag{46}$$

PROOF. First, a notation to make calculations easier, let $\alpha = 1 - \frac{1}{c}$. Then we have:

$$\begin{cases} \Phi_{m,k} = 1 - \frac{a(k-1)+\frac{1}{c}}{m+k-1} \\ \Psi_{m,k} = 1 + \frac{a(k-1)}{m} \end{cases}$$

The middle term can be written using $2 + \frac{1}{c} = c$,

$$\begin{aligned} \frac{1}{c} + \Psi_{m,k} + \Phi_{m,k+1} &= 2 + \frac{1}{c} + \frac{a(k-1)}{m} - \frac{ak + \frac{1}{c}}{m+k} \\ &= c + \frac{a(k-1)}{m} - \frac{ak + \frac{1}{c}}{m+k} \quad . \end{aligned}$$

Let's start by proving the right inequality.

1. **Right inequality:** The right side can be written:

$$c\Phi_{m+1,k}\Psi_{m,k} = c + c\left[\frac{ak}{m} - \frac{a(k-1)+\frac{1}{c}}{m+k} - \frac{a(k-1)}{m}\frac{\left(a(k-1)+\frac{1}{c}\right)}{m+k}\right] \tag{47}$$

The inequality we want to prove is equivalent to, dropping the first c: For all $m, k \geq 1$:

$$\begin{aligned} &\frac{a(k-1)}{m} - \frac{ak+\frac{1}{c}}{m+k} \leq c\left[\frac{a(k-1)}{m} - \frac{a(k-1)+\frac{1}{c}}{m+k} - \frac{a(k-1)}{m}\frac{\left(a(k-1)+\frac{1}{c}\right)}{m+k}\right] \\ &\Leftrightarrow a(k-1)(m+k) - m(ak + \frac{1}{c}) \leq c\left[a(k-1)(m+k) - m\left(a(k-1)+\frac{1}{c}\right) - a(k-1)\left(a(k-1)+\frac{1}{c}\right)\right] \\ &\Leftrightarrow akm + ak^2 - am - ak - mak - \frac{m}{c} \leq c\left[akm + ak^2 - ma - ak - akm + ma - \frac{m}{c} - a^2(k-1)^2 - \frac{a}{c}k + \frac{a}{c}\right] \\ &\Leftrightarrow a\left(c - ac - 1\right)k^2 + ac\left(2a - 1\right)k + m\left(a + \frac{1}{c} - 1\right) + a - a^2c \geq 0 \end{aligned}$$

However, $c - ac - 1 = 0$ and $a + \frac{1}{c} - 1 = 0$. Thus, the left side above gives rise to an affine function f in k defined as: $f(k) = ac(2a-1)k + a - a^2c$ that verifies $f(1) = ac(2a-1) + a - a^2c = 0$, and since its slope $ac(2a-1) = a(2c-3) = a(\sqrt{2}-1) > 0$, we have $f(k) \geq 0, \quad \forall k \geq 1$. Therefore, since all inductions above are equivalent to each other, the right inequality is proven.

2. **Left inequality:** Similarly, the left side can be written:

$$c\Phi_{m,k+1}\Psi_{m,k+1} = c + c\left[\frac{ak}{m} - \frac{ak+\frac{1}{c}}{m+k} - \frac{ak}{m}\frac{\left(ak+\frac{1}{c}\right)}{m+k}\right] \tag{48}$$

Again c cancels out, and the inequality is equivalent to, for all $m, k \geq 1$:

$$\begin{aligned}
&\frac{a(k-1)}{m} - \frac{ak+\frac{1}{c}}{m+k} \geq c\left[\frac{ak}{m} - \frac{ak+\frac{1}{c}}{m+k} - \frac{ak}{m}\frac{\left(ak+\frac{1}{c}\right)}{m+k}\right] \\
&\Leftrightarrow akm + ak^2 - am - ak - mak - \frac{m}{c} \geq c\left[akm + ak^2 - akm - \frac{m}{c} - a^2k^2 - \frac{ak}{c}\right] \\
&\Leftrightarrow a\left(c - ac - 1\right)k^2 + m\left(a + \frac{1}{c} - 1\right) \geq 0
\end{aligned}$$

However, $c - ac - 1 = 0$ and $a + \frac{1}{c} - 1 = 0$. Thus, we indeed have the last inequality. Therefore, since all inductions above are equivalent to each other, the left inequality is proven. □

Proof of proposition 3 We can now describe our induction proof. We would like to show for all m the statement: $P(m) : (\forall k \geq 1)\ A(m,k)$ and $B(m,k)$.

0. intialization step For $m = 1$, on one hand we have for all $k \geq 1 : D_{1,k} = 1$ and $c\Phi_{1,k} = 1 + \frac{c-2}{k} = 1 + \frac{\sqrt{2}-1}{k} \geq 1$, thus we have $A(1,k)\ \forall k$. On the other hand, one can easily show that $D_{2,1+k} = 2k+1$ and that $c\Psi_{1,k} = (c-1)k + 1 = \sqrt{2}k + 1 \leq 2k+1$, since $D_{1,k+1} = 1$, we have $B(1,k)\ \forall k$.

1. induction step (on m) . Let $m \geq 2$ and assume $A(m,k)$ and $B(m,k)$ are true for all $k \geq 1$. We first start by proving $A(m+1,k)$ for any $k \geq 1$.

1.1 $A(m,k)$ **and** $B(m,k)\,(\forall k) \Rightarrow A(m+1,k)\,(\forall k)$**:** We show this directly for any $k \geq 1$. Using the recursive definition of Delannoy numbers (38) applied to left side of $A(m+1,k)$ we have:

$$D_{m+1,m+k+1} = D_{m+1,m+k} + D_{m,m+k+1} + D_{m,m+k} \ . \tag{49}$$

Applying $A(m,k+1)$ to the second term of the right side we get: $D_{m,m+k+1} \leq c\Phi_{m,k+1}D_{m,m+k}$; and applying $B(m,k)$ to the third term, we get: $D_{m,m+k} \leq \frac{D_{m+1,m+k}}{c\Psi_{m,k}}$. Which sums up to: $D_{m+1,m+k+1} \leq \left(1 + \frac{1}{c\Psi_{m,k}} + \frac{\Phi_{m,k+1}}{\Psi_{m,k}}\right)D_{m+1,m+k}$. To conclude $A(m+1,k)$, all we need is:

$$\left(1 + \frac{1}{c\Psi_{m,k}} + \frac{\Phi_{m,k+1}}{\Psi_{m,k}}\right) \leq c\Phi_{m+1,k} \ , \tag{50}$$

which follows directly from the right inequality of Lemma A.2. We have thus proven $A(m+1,k)$ for any arbitrary $k \geq 1$.

1.2 $A(m,k)$**,** $A(m+1,k)$**,** $B(m,k)\,(\forall k) \Rightarrow B(m+1,k)\,(\forall k)$**:** We prove the statement $B(m+1,k)\,(\forall k)$ via an induction reasoning on k.

1.2.0 initialization step (k = 1) . For $k = 1$, we have to show that $c\Psi_{m+1,1}D_{m+1,m+2} \leq D_{m+2}$. On one hand, we have $\Psi_{m+1,1} = 1$. On the other hand, using the recursion definition (37) we get: $D_{m+2} = D_{m+1,m+2} + D_{m+2,m+1} + D_{m+1}$. And by symmetry of Delannoy numbers: $D_{m+2} = 2\ D_{m+1,m+2} + D_{m+1}$. Now using the growth lemma A.1 on D_{m+1}, we have: $D_{m+1,m+2} \leq \frac{c^2-1}{2c^2}D_{m+2}$. Since $c = 1+\sqrt{2}$, we have $\frac{c^2-1}{2c^2} = \frac{1}{c}$ which concludes $B(m+1,1)$.

1.2.1 induction step (on k) :. Let $k \geq 1$ and assume $B(m+1,k)$ is true, let's prove that $B(m+1,k+1)$ is true as well. $B(m+1,k+1)$ can be written: $c\Psi_{m+1,k+1}D_{m+1,m+k+2} \leq D_{m+2,m+k+2}$. Again, using the recursion definition, we have:

$$D_{m+2,m+k+2} = D_{m+1,m+k+2} + D_{m,m+k+1} + D_{m+1,m+k+1} + D_{m+2,m+k+1} \tag{51}$$

Applying the already proven $A(m+1,\ k')$ (for all k') to the second member of the right side, we have: $D_{m+1,m+k+1} \geq \frac{D_{m+1,m+k+2}}{c\Phi_{m+1,k+1}}$. Similarly, applying the induction (on k) assumption $B(m+1,\ k)$ to the third member, we get: $D_{m+2,m+k+1} \geq c\Psi_{m+1,k}D_{m+1,m+k+1}$. Which sums up to: $D_{m+2,m+k+2} \geq \left(1 + \frac{1}{c\Phi_{m+1,k+1}} + \frac{\Psi_{m+1,k}}{\Phi_{m+1,k+1}}\right)D_{m+1,m+k+2}$. To conclude $B(m+1,k+1)$, all we need is:

$$c\Psi_{m+1,k+1} \leq \left(1 + \frac{1}{c\Phi_{m+1,k+1}} + \frac{\Psi_{m+1,k}}{\Phi_{m+1,k+1}}\right) \tag{52}$$

Which follows directly from the left inequality of Lemma A.2, where m is substituted with $m+1$. Therefore, $B(m+1, k+1)$ is true, ending the induction proof on k.

Hence, $B(m+1, k)$ holds for any $k \geq 1$, the induction on proof on m is complete. □

A.3 Other proofs

Lemma A.3 (Bounded transported mass)**.** *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^p$ *and* $\mathbf{P}_{\mathbf{x},\mathbf{y}} \in \mathbb{R}_+^{p\times p}$ *their associated transport plan, solution of* (2)*. Let* $\kappa = \min_{i,j} e^{-\frac{M_{ij}}{\gamma}}$*. We have the following bounds on the total transported mass:*

$$\kappa\|\mathbf{x}\|_1\|\mathbf{y}\|_1 \leq \|\mathbf{P}_{\mathbf{x},\mathbf{y}}\|_1^{2+\frac{\varepsilon}{\gamma}} \leq p^{2(1+\frac{\varepsilon}{\gamma})}\|\mathbf{x}\|_1\|\mathbf{y}\|_1 \tag{53}$$

PROOF. The first order optimality condition of (2) reads for all $i, j \in [\![1, p]\!]$:

$$\varepsilon \log(\mathbf{P}_{ij}) - \varepsilon \log(\mathbf{K}_{ij}) + \gamma \log\left(\frac{\mathbf{P}_{i.}^\top \mathbb{1} \mathbf{P}_{.j}^\top \mathbb{1}}{\mathbf{x}_i \mathbf{y}_j}\right) = 0 \tag{54}$$

$$\Leftrightarrow \mathbf{P}_{ij}^{\frac{\varepsilon}{\gamma}} \mathbf{P}_{i.}^\top \mathbb{1} \mathbf{P}_{.j}^\top \mathbb{1} = \mathbf{x}_i \mathbf{y}_j e^{-\frac{\mathbf{M}_{ij}}{\gamma}} \tag{55}$$

On one hand we have:

$$\begin{aligned}\sum_{i,j}^{p} \mathbf{P}_{ij}^{\frac{\varepsilon}{\gamma}} \mathbf{P}_{i.}^\top \mathbb{1} \mathbf{P}_{.j}^\top \mathbb{1} &\leq \|\mathbf{P}\|_\infty^{\frac{\varepsilon}{\gamma}} \sum_{i,j}^{p} \mathbf{P}_{i.}^\top \mathbb{1} \mathbf{P}_{.j}^\top \mathbb{1} \\ &= \|\mathbf{P}\|_\infty^{\frac{\varepsilon}{\gamma}} \|\mathbf{P}\|_1^2 \\ &\leq \|\mathbf{P}\|_1^{2+\frac{\varepsilon}{\gamma}}\end{aligned}$$

On the other hand, using Jensen's inequality in the second step:

$$\begin{aligned}\sum_{i,j}^{p} \mathbf{P}_{ij}^{\frac{\varepsilon}{\gamma}} \mathbf{P}_{i.}^\top \mathbb{1} \mathbf{P}_{.j}^\top \mathbb{1} &\geq \sum_{i,j}^{p} \mathbf{P}_{ij}^{\frac{\varepsilon}{\gamma}+2} \\ &\geq p^2 \left(\frac{\sum_{i,j} \mathbf{P}_{ij}}{p^2}\right)^{\frac{\varepsilon}{\gamma}+2} \\ &\geq p^{-2-2\frac{\varepsilon}{\gamma}} \|\mathbf{P}\|_1^{2+\frac{\varepsilon}{\gamma}}\end{aligned}$$

Finally, since $\kappa \leq \min_{ij} e^{-\frac{\mathbf{M}_{ij}}{\gamma}} \leq 1$ and $\frac{2+2\frac{\varepsilon}{\gamma}}{2+\frac{\varepsilon}{\gamma}} \leq \frac{3}{2}$,we get the desired inequalities. □

Differentiability of W

Proposition A.4. *Given a fixed* $\mathbf{y} \in \mathbb{R}_+^{n\times p}$*, the unbalanced Wasserstein distance function* $\mathbf{x} \to W(\mathbf{x}, \mathbf{y})$ *is smooth and its gradient is given by:*

$$\nabla_{\mathbf{x}} W(\mathbf{x}, \mathbf{y}) = \gamma(1 - a(\mathbf{x}, \mathbf{y})^{-\frac{\varepsilon}{\gamma}}) \tag{56}$$

Where $\mathbf{a}(\mathbf{x}, \mathbf{y})$ *is the optimal Sinkhorn scaling, solution of the fixed point problem* (57)*.*

PROOF. As noted by Feydy et al. (2017). The proof is similar to the balanced case. Indeed by applying the envelope theorem to the equivalent dual problem (3), one has $\nabla_{\mathbf{x}} W(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{x}}\left(-\gamma\langle e^{-\frac{u}{\gamma}} - 1, \mathbf{x}\rangle\right) = \gamma\left(1 - e^{-\frac{u}{\gamma}}\right) = \gamma(1 - a(\mathbf{x}, \mathbf{y})^{-\frac{\varepsilon}{\gamma}})$ with the change of variable $\varepsilon \log(a) = u$. □

Proposition A.5. *Let* $\mathbf{y}, \mathbf{x} \in \mathbb{R}_{++}^{n\times p}$ *be a stationary point of* S *i.e* $\nabla S(\mathbf{x}, \mathbf{y}) = (\mathbf{0}, \mathbf{0})$*. Then,* $S(\mathbf{x}, \mathbf{y}) = 0$*. Moreover, if* $\mathbf{K}$ *is positive definite, then* $\mathbf{x} = \mathbf{y}$*.*

PROOF. Let $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ the solutions of the fixed problems:

$$\mathbf{a} = \left(\frac{\mathbf{x}}{K\mathbf{b}}\right)^{\omega} \quad , \quad \mathbf{b} = \left(\frac{\mathbf{y}}{K\mathbf{a}}\right)^{\omega}, \quad \mathbf{c} = \left(\frac{\mathbf{x}}{K\mathbf{c}}\right)^{\omega}, \quad \mathbf{d} = \left(\frac{\mathbf{y}}{K\mathbf{d}}\right)^{\omega} \tag{57}$$

We have applying the chain rule, $\frac{1}{2}$ disappears and we get: $\nabla_x S(\mathbf{x}, \mathbf{y}) = \gamma(\mathbf{c}^{-\frac{\varepsilon}{\gamma}} - \mathbf{a}^{-\frac{\varepsilon}{\gamma}})$ and $\nabla_y S(\mathbf{x}, \mathbf{y}) = \gamma(\mathbf{d}^{-\frac{\varepsilon}{\gamma}} - \mathbf{b}^{-\frac{\varepsilon}{\gamma}})$

If $(\mathbf{x}, \mathbf{y})$ is a stationary point of S, then we immediately have $\mathbf{a} = \mathbf{c}$ and $\mathbf{b} = \mathbf{d}$. The fixed point equations lead to $\mathbf{Ka} = \mathbf{Kb} = \mathbf{Kc} = \mathbf{Kd}$.

The transported mass between $\mathbf{x}$ and $\mathbf{y}$ is given by: $\|P_{\mathbf{x},\mathbf{y}}\|_1 = \langle \mathbf{a}, \mathbf{Kb} \rangle = \langle \mathbf{b}, \mathbf{Ka} \rangle$. Therefore, using Proposition 6, S can be written:

$$\begin{aligned} S(\mathbf{x}, \mathbf{y}) &= \frac{\varepsilon + 2\gamma}{2}(\langle \mathbf{c}, \mathbf{Kc} \rangle + \langle \mathbf{d}, \mathbf{Kd} \rangle - 2\langle \mathbf{a}, \mathbf{Kb} \rangle) \\ &= \frac{\varepsilon + 2\gamma}{2}(\langle \mathbf{c}, \mathbf{Kc} \rangle + \langle \mathbf{d}, \mathbf{Kd} \rangle - 2\langle \mathbf{a}, \mathbf{Kb} \rangle - \langle \mathbf{b}, \mathbf{Ka} \rangle) \\ &= \frac{\varepsilon + 2\gamma}{2}(\langle \mathbf{c} + \mathbf{d} - \mathbf{a} - \mathbf{b}, \mathbf{Ka} \rangle) \\ &= 0 \end{aligned}$$

Moreover, if $\mathbf{K}$ is positive definite, $\mathbf{Ka} = \mathbf{Kb}$ leads to $\mathbf{a} = \mathbf{b}$ and thus $\mathbf{x} = \mathbf{y}$. □

B Experiments

In this section, we provide details on the experimental settings as well as complementary results.

B.1 Temporal shift bound

We proved in the theoretical section that our quadratic bound holds for $0 < \beta \leq \frac{\mu}{\log(3TD_{T,T})}$. We argue that this upper bound is too restrictive in practice. We show the empirical comparisons for different values of beta. When β is too large, (here $\beta > 10$) one can see that the quadratic bound does not hold for large temporal shifts. To get a tighter general bound, one must carry out a finer analysis of the remaining logsumexp terms after isolating the first large term $\frac{n_0}{n_0'}$.

B.2 Brain imaging

The time series realizations are defined on the surface of the brain which is modeled as a triangulated mesh of 642 vertices. We compute the squared ground metric M on the the mesh using Floyd-Warshall's algorithm and normalize it by its median. This normalization is standard in several optimal transport applications (Peyré and Cuturi, 2018) and allows to scale the entropy hyperparameter ε to the dimension of the data. We set $\varepsilon = 10/642$ and $\gamma = 1$ given by the heuristic proposed by Janati et al. (2019). Figure 10 shows additional t-SNE embeddings with different values of β.

B.3 handwritten letters

The raw handwritten letters data consist of (x, y) coordinates of the trajectory of the pen. The data include between 100 and 205 strokes – time point – for each sample. The preprocessing we performed consisted of creating the images of the cumulated trajectories and rescaling them in order to fit into a (64 × 64) 2D grid. Smoothing the data spatially. Figure 12 shows more examples of the processed data. The time series realizations are defined on a 2D grid We compute the squared ground metric M on the the mesh using Floyd-Warshall's algorithm and normalize it by its median. This normalization is standard in several optimal transport applications (Peyré and Cuturi, 2018) and allows to scale the entropy hyperparameter ε to the dimension of the data. We set $\varepsilon = 10/642$ and $\gamma = 1$ given by the heuristic proposed by Janati et al. (2019). Figure 11 shows additional t-SNE embeddings with different values of β.

C Code

The code is provided in the supplementary materials folder. Please follow the guidelines in the README file to reproduce all experiments.

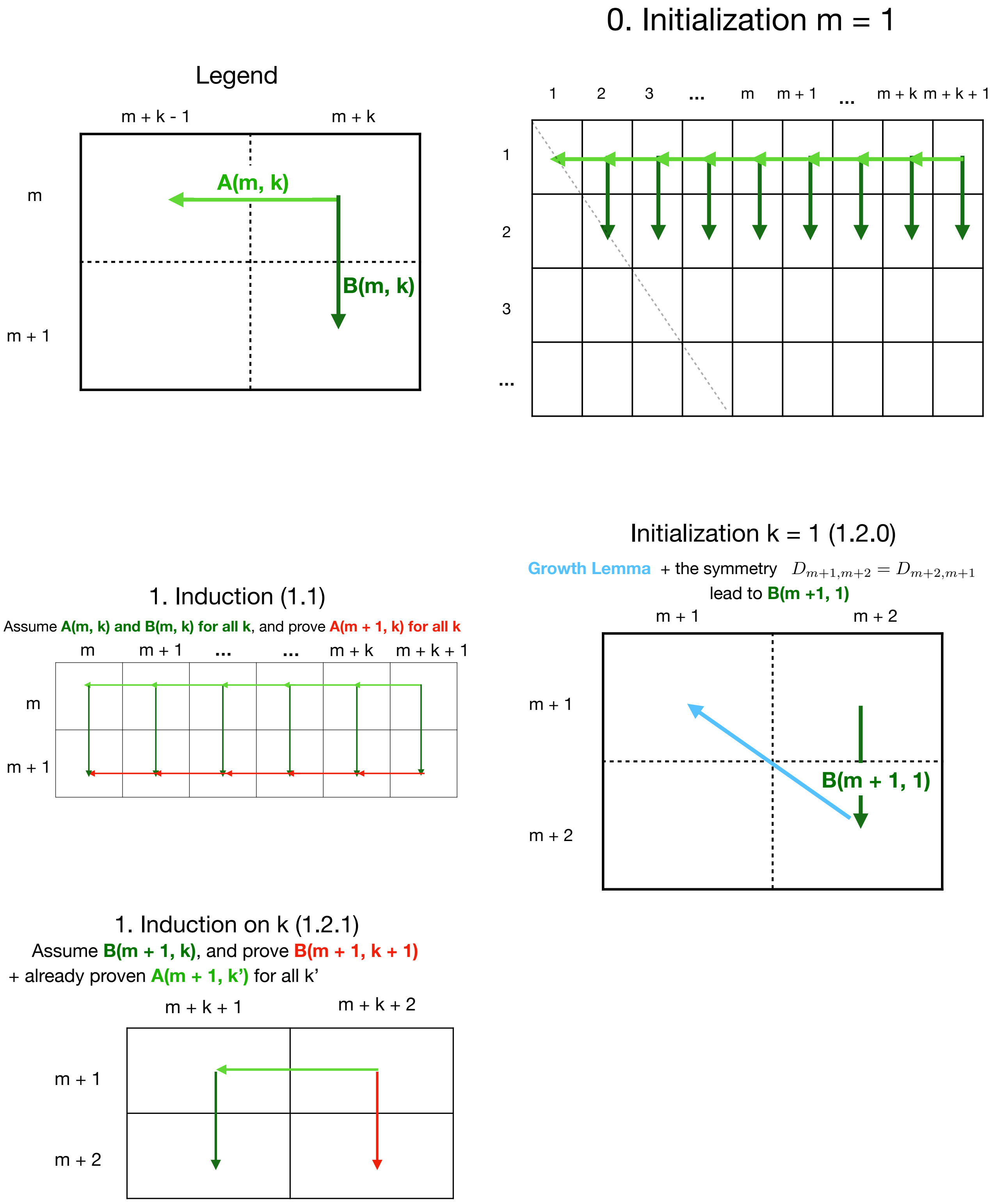

Legend
m + k - 1
m + k
m
m + 1
A(m, k)
B(m, k)
0. Initialization m = 1
1
2
3
...
m
m + 1
...
m + k
m + k + 1
1
2
3
...
1. Induction (1.1)
Assume A(m, k) and B(m, k) for all k, and prove A(m + 1, k) for all k
m
m + 1
...
...
m + k
m + k + 1
m
m + 1
Initialization k = 1 (1.2.0)
Growth Lemma + the symmetry $D_{m+1,m+2} = D_{m+2,m+1}$
lead to B(m +1, 1)
m + 1
m + 2
m + 1
m + 2
B(m + 1, 1)
1. Induction on k (1.2.1)
Assume B(m + 1, k), and prove B(m + 1, k + 1)
+ already proven A(m + 1, k') for all k'
m + k + 1
m + k + 2
m + 1
m + 2

Fig. 8. Visualization of the proof of proposition 3. The key steps are 1.1 and 1.2.1, where given the top and left arrows, one must derive the right and bottom arrows.

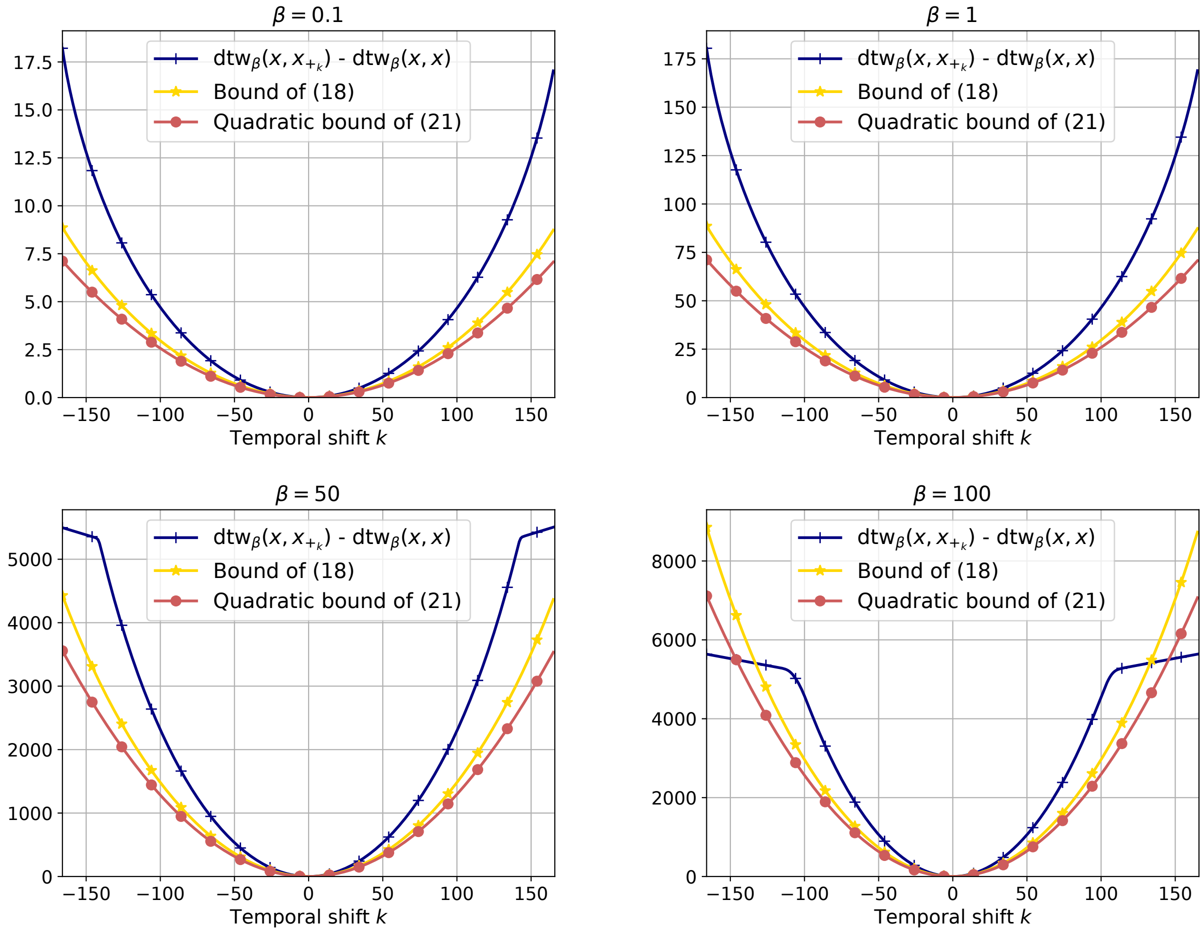


Fig. 9. Empirical evaluation of the obtained theoretical bounds for various values of β

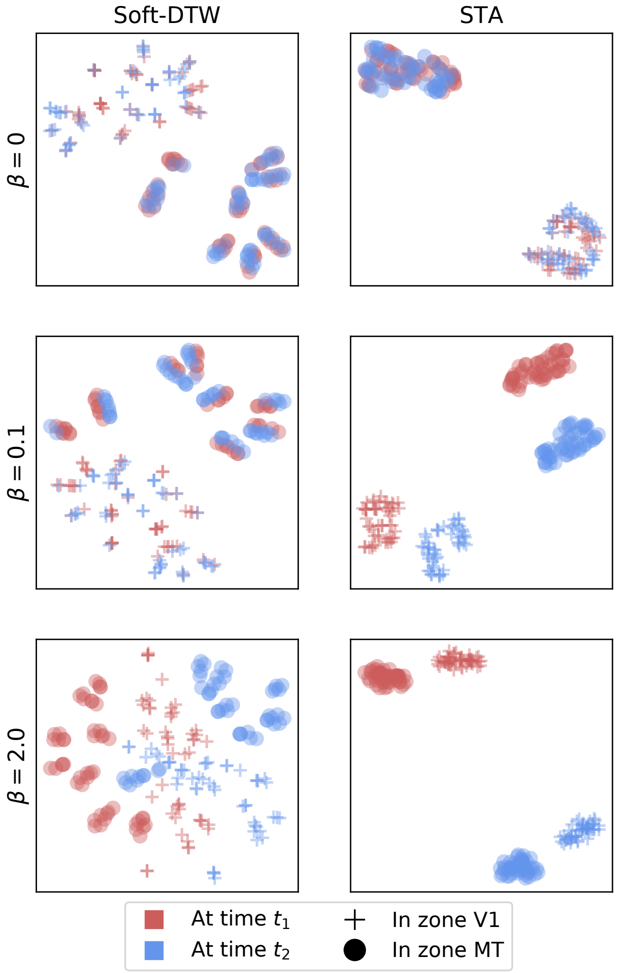


Fig. 10. tSNE embeddings of the data. STA (proposed) captures spatial variability. Increasing β helps capture more temporal variability.

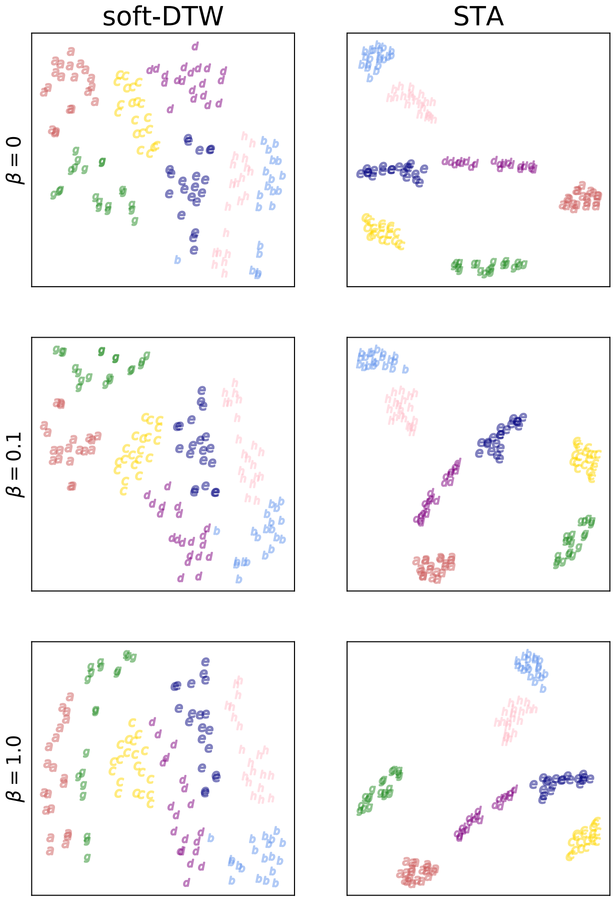


Fig. 11. tSNE embeddings of the data. STA (proposed) captures spatial variability.

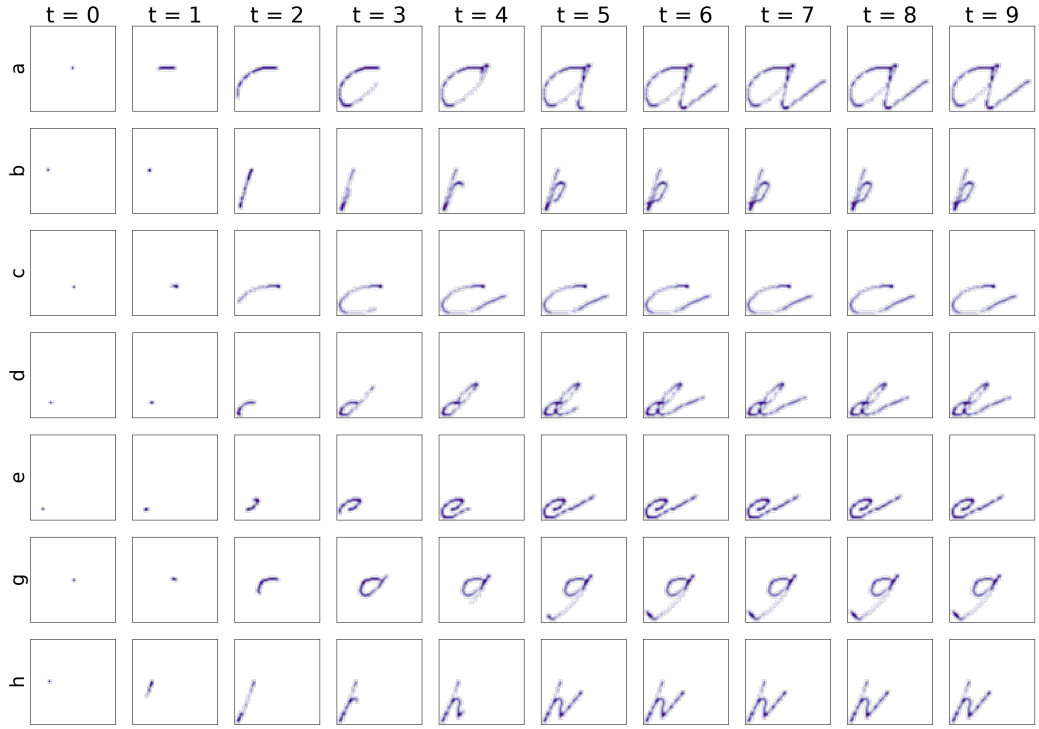


Fig. 12. An example of each handwritten letter in the dataset