Journal of Machine Learning Research 23 (2022) 1-30 Submitted 3/22; Revised 10/22; Published 02/25

Averaging Spatio-temporal Signals using Optimal Transport and Soft Alignments

Hicham Janati HJANATI@INSEA.AC.MA
INSEA, Morocco

Marco Cuturi MARCOCUTURI@ENSAE.FR
CREST ENSAE
Paris, France

Alexandre Gramfort ALEXANDRE.GRAMFORT@INRIA.FR
Université Paris-Saclay, Inria, CEA, Palaiseau, France

Editor: Joan Bruna

Abstract

Several fields in science, from genomics to neuroimaging, require monitoring populations (measures) that evolve with time. These complex datasets, describing dynamics with both time and spatial components, pose new challenges for data analysis. We propose in this work a new framework to carry out *averaging* of these datasets, with the goal of synthesizing a representative template trajectory from multiple trajectories. We show that this requires addressing three sources of invariance: shifts in time, space, and total population size (or mass/amplitude). Here we draw inspiration from dynamic time warping (DTW), optimal transport (OT) theory and its unbalanced extension (UOT) to propose a criterion that can address all three issues. This proposal leverages a smooth formulation of DTW (Soft-DTW) that is shown to capture temporal shifts, and UOT to handle both variations in space and size. Our proposed loss can be used to define spatio-temporal barycenters as Fréchet means. Using Fenchel duality, we show how these barycenters can be computed efficiently, in parallel, via a novel variant of entropy-regularized debiased UOT. Experiments on handwritten letters and brain imaging data confirm our theoretical findings and illustrate the effectiveness of the proposed loss for spatio-temporal data.

Keywords: Time-Series, Dynamic time warping, Optimal transport

1. Introduction

One of the most elementary operations in machine learning pipelines is to summarize data by aggregating several samples into one. On metric spaces, while the Euclidean mean is probably the most standard averaging tool, it is often not suited to applications where features are inherently structured independently of the data. This is for instance the case with spatio-temporal data where features correspond to fixed physical positions in space in some given moments in time. Spatio-temporal data can be seen as multivariate time-series where each (temporal) observation is defined in a space equipped with a physical notion of geometry. For instance, with videos displaying the motion of some object, the spatial features (pixels) are defined through a fixed rectangular grid while the temporal features are given in a chronological order. More generally, the spatial geometry can be defined on a predefined graph. This is the case with brain imaging data, where the measurements of

neural activity are provided at each vertex of the triangulated mesh of the brain and across several time points (Gramfort et al., 2011).

Formally, each sample μ^i in our dataset can be represented as a time sequence of discrete non-negative measures $(\mu_t^i)_{t=1..T}$ defined in some finite set $\mathcal{X}$. Our aim is to find a loss function $\mathcal{L}$ through which the weighted Fréchet mean can be defined as:

$$\bar{\mu} = \underset{\mu \in (\mathcal{M}_+(\mathcal{X}))^T}{\arg\min} \sum_{i=1}^{N} w_i \mathcal{L}(\mu^i, \mu) \ , \tag{1}$$

where $(w_i)i = 1..N$ is a set of positive weights summing to one.

In the videos example, $\mathcal{X} \overset{\text{def}}{=} \{x_1, \dots, x_p\}$ would be the set of pixels and $\mu_t^i \in \mathcal{M}_+(\mathcal{X})$ would correspond to the t-th frame of the i-th video. Given that μ_t^i is discrete with a fixed support $\mathcal{X}$, it can be identified with a non-negative vector of weights $\mathbf{p} \in \mathbb{R}_+^p$ such that $\mu_t^i = \sum_{k=1}^p \mathbf{p}_k \delta_{x_k}$. Such a setting was considered in Janati et al.'s proposal 2020a where they defined $\mathcal{L}$ in (1) as an alignment soft-DTW metric operating on top of an OT cost. Intuitively, DTW is computed by finding the alignment between time points that has the lowest cost. If this cost is defined through OT, then the obtained optimal temporal alignment would pair time points that are *spatially* similar to each other while capturing temporal differences. Janati et al. (2020a) propose to use Soft-DTW (Cuturi and Blondel, 2017) – rather than DTW (Sakoe and Chiba, 1978) – to benefit from additional properties notably a different behaviour with respect to time shifts (Janati et al., 2020a). In this paper, our work sees instead in Soft-DTW a differentiable loss that will be more adequate for the task of computing barycenters.

Related work On one hand, ignoring the temporal dimension, one could leverage the geometric properties of optimal transport (OT) metrics by computing OT barycenters (i.e using an OT loss for $\mathcal{L}$) sequentially through time i.e averaging all the samples for each time point t independently of the others. These geometrical differences between measures are captured through some pre-defined ground metric between the spatial features. Taking once again the videos example, this ground metric can be for instance given by the Euclidean distance between the coordinates of the pixels in their fixed rectangular frame. A straightforward use of OT with spatio-temporal data is to consider the transportation across space and time simultaneously via a customized ground metric. Motivated by OT between signed signals, this customization of the ground metric was previously studied by Thorpe et al. (2017) and introduced as TL_p distances. However, this method ignores the chronology of the data and requires a difficult tuning of the tradeoff between spatial and temporal transport costs when designing the ground metric.

Recently, Vayer et al. (2020) proposed a new derivative of soft-DTW that learns the global invariances of the data that is very well adapted for data depicting trajectories that are invariant by rotations. This method however is not suited for comparing and averaging complex mass dynamics that cannot be modeled as a set of trajectories which is the case with meteorological or neuroimaging data for instance. Inspired by the Gromov-Wasserstein framework, this proposal was later generalized by Cohen et al. (2021) to align and average time series in different spaces.

Challenges and contributions Following Janati et al. (2020a), we propose to define the loss function $\mathcal{L}$ using Soft-DTW and unbalanced optimal transport (UOT). This combination

is straightforward since Soft-DTW requires a cost function of its own to align temporal observations across time series. Using UOT as a cost function leads to the loss function coined STA: *spatio-temporal alignement*. While evaluating STA can be easily carried out to perform clustering or metric based classification, computing its Fréchet mean comes with several challenges:

- A classic bottleneck when using OT is scalability. In Janati et al. (2020a), the differentiability of Soft-DTW is more or less hinted, but not experimented with. Attempting to solve (1) naively via gradient descent turns out to be computationally infeasible as one gradient step requires computing NT^2 gradients of entropic OT, and thus running NT^2 Sinkhorn loops. Instead we propose to re-write the loss function using Fenchel duality so that one can update the spatio-temporal barycenter using the generalized Sinkhorn's algorithm for barycenters. Using Fenchel duality is the main contribution of this paper. It transforms a difficult spatio-temporal barycenter problem to a straightforward spatial barycenter weighted in time. Interestingly, this result is not specific to optimal transport: spatio-temporal barycenters for any spatial loss function can be computed via alternating between temporal weight updates and purely spatial averages.

- As in (Janati et al., 2020a), we use a debiased variant of unbalanced OT that is non-negative and suffers from less entropic blur. Starting from Janati et al.'s proposal 2020b to modify the Sinkhorn algorithm to compute *debiased* OT barycenters (that cancel the blurring effect induced by entropic regularization), we extend their work and generalize it to the *unbalanced* case, where the marginals have different sums.

- As is usual in mixed approaches that combine several metrics (Thorpe et al., 2017; Damodaran et al., 2018), setting hyperparameters that control the tradeoffs between these different contributions is challenging. We propose a heuristic to set the hyperparameters of OT and soft-DTW based on a maximum desired temporal shift.

- To find this heuristic, we derive tight inequalities that bound the growth of Delannoy numbers which can be of independent interest in Combinatorics.

Structure Section 2 provides background material on Soft-DTW, our contributions to Delannoy numbers that lead to a practical heuristic to set the Soft-DTW hyperparameter. Next in section 3, we propose an alternating optimization algorithm to compute the Fréchet mean (1). This alternating algorithm reduces the problem to a sequence of temporally weighted Fréchet means of the Soft-DTW cost. In section 4, we propose a generalized Sinkhorn algorithm to compute these Fréchet means for debiased unbalanced OT barycenters. Finally, we showcase the performance of STA averaging in Section 5 in a forecasting experiment of handwritten characters and averaging of brain imaging data.

Notation Vectors are denoted by small cap boldface letters (e.g $\mathbf{x}$), matrices are denoted by large cap boldface letters (e.g $\mathbf{X}$). For a matrix $\mathbf{X}$, rows and columns are denoted by $\mathbf{X}_{i.}$ and $\mathbf{X}_{.i}$. Some operations on matrices are considered element-wise: log, exp, and division. The element-wise product is denoted by $\odot$. The entropy of a vector (or a matrix) $\mathbf{x} \in \mathbb{R}^p$ is defined as $H(\mathbf{x}) = \langle \mathbf{x}, \log(\mathbf{x}) - \mathbb{1}_p \rangle$. For any integer $q \in \mathbb{N}$, $[\![q]\!]$ is the set $\{1, \dots, q\}$.

The set of vectors in $\mathbb{R}^p$ with non-negative (resp. positive) entries is denoted by $\mathbb{R}^p_+$ (resp. $\mathbb{R}^p_{++}$). The Kullback-Leibler (KL) divergence between two positive vectors (or matrices) by $\mathrm{KL}(\mathbf{x}|\mathbf{y}) = \langle \mathbf{x}, \log(\mathbf{x}/\mathbf{y})\rangle + \langle \mathbf{y} - \mathbf{x}, \mathbb{1}_p\rangle$ with the continuous extensions $0\log(0/0) = 0$ and $0\log(0) = 0$ and the limit $\mathbf{x} \neq 0 \Rightarrow \mathrm{KL}(\mathbf{x}|0) = +\infty$. The Delannoy set $\mathcal{A}_{T_1,T_2}$. corresponds to the feasible set of binary matrices of $\mathbb{R}^{T_1\times T_2}$ where only $\rightarrow, \downarrow, \searrow$ movements are allowed. These binary matrices form a non-zero path starting at (1,1) and ending at (T_1, T_2), thus the condition $A_{1,1} = A_{T_1,T_2} = 1$ holds for any alignment matrix $A \in \mathcal{A}$.

2. Soft-DTW and Delannoy numbers

2.1 Background

Dynamic time warping Consider two multivariate time series $\mathbf{x} \in \mathbb{R}^{p,T_1}$ and $\mathbf{y} \in \mathbb{R}^{p,T_2}$ with respective lengths T_1, T_2 and having observations in $\mathbb{R}^p$. DTW is defined through some pairwise distance matrix $\Delta(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{T_1,T_2}$ between all their time points such that the cost of a given alignment function $\sigma : [\![1, T_1]\!] \to [\![1, T_2]\!]$ is equal to $\sum_{i=1}^{T_1} \delta(\mathbf{x}_i, \mathbf{y}_{\sigma(i)})$, where we used an upper case Δ to differentiate the distance matrix from the distance function δ it was defined through i.e $\Delta_{ij} = \delta(\mathbf{x}_i, \mathbf{y}_j)$ To guarantee the preservation of the chronology of the data, σ must be increasing and verify $\sigma(1) = 1$ and $\sigma(T_1) = T_2$. The resulting optimization problem is however better posed as a minimization of $\sum_{i=1}^{T_1}\sum_{j=1}^{T_2} A_{ij}\delta(\mathbf{x}_i, \mathbf{y}_j)$ over the set of binary alignments A on the rectangular lattice $[\![1, T_1]\!] \times [\![1, T_2]\!]$ where no temporal back steps are allowed. This amounts to considering binary matrices with a non-zero path linking the corners of the lattice $(1, 1)$ (upper left) and (T_1, T_2) (bottom right) using $\rightarrow, \downarrow, \searrow$ steps exclusively (Sakoe and Chiba, 1978). Formally, DTW is defined as:

$$\mathbf{dtw}(\mathbf{x}, \mathbf{y}; \Delta) = \min\{\langle \mathbf{A}, \Delta(\mathbf{x}, \mathbf{y})\rangle, \mathbf{A} \in \mathcal{A}_{T_1,T_2}\} \ , \tag{2}$$

where $\langle ., .\rangle$ denotes the Frobenius dot product. The binary nature of the constraint set in (2) 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 regularized variants of DTW (Saigo et al., 2004; Cuturi et al., 2007; Cuturi, 2011; Cuturi and Blondel, 2017). Instead of selecting *the* minimum cost alignment, Global Alignment Kernels (GAK) for instance (Saigo et al., 2004; Cuturi et al., 2007; Cuturi, 2011) compute a weighted cost of all possible alignments with a certain smoothing hyperparameter. Similarly, the soft-minimum generalization approach of Cuturi and Blondel (2017) – called soft-DTW – provides a similar framework to that of GAK that includes DTW as a sub-case:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}; \Delta) = \mathrm{softmin}_\beta\{\langle \mathbf{A}, \Delta(\mathbf{x}, \mathbf{y})\rangle, \mathbf{A} \in \mathcal{A}_{T_1,T_2}\} \ , \tag{3}$$

where the soft-minimum operator of a set $\mathcal{A}$ with parameter $\beta \geq 0$ is defined as:

$$\mathrm{softmin}_\beta(\mathcal{A}) = \begin{cases} -\beta \log\left(\sum_{a\in\mathcal{A}} e^{-a/\beta}\right) & \text{if } \beta > 0 \\ \min \mathcal{A} \quad \text{if } \beta = 0 \end{cases} . \tag{4}$$

In particular, softmin is continuous at 0 so that when $\beta \to 0$, $\mathbf{dtw}_\beta$ approaches $\mathbf{dtw}$.

Remark 1 *To control for time series lengths,* $\mathbf{dtw}_0$ *can be normalized by the total length,* $1/(T_1 + T_2)$*, as proposed in (Sakoe and Chiba, 1978, Eq.(22)). This is particularly important*

for datasets with time series with very different sizes. For the barycenter problem, this normalization can be included in the weights vector w defined in (1). We did not explore this normalization given that the datasets we considered in the experimental section were homogeneous in length.

Forward recursion 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.

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

Input: data $\mathbf{x}, \mathbf{y}$ soft-min parameter β and distance function δ
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**
    $r_{i,j} = \delta(\mathbf{x}_i, \mathbf{y}_j) + \mathrm{softmin}_\beta(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1})$
  end for
end for

Remark 2 *The boundary conditions hard-coded by the Delannoy set do not allow for alignments that disregard the starting and ending point of the time series. This is enforced in Algorithm 1 by setting all the initial costs of the first row and column of r to $+\infty$. However, by changing the initialization of r, it is possible to relax these constraints. In situations where partial matching of time series is relevant, one can set the first row to the costs $r_{0,j} = \delta(x_0, y_j)$ and the first column to $r_{i,0} = \sum_{k=0}^{i} \delta(x_k, y_0)$. The final cost is obtained by picking the minimizer of the last row.*

Algorithmic differentiation When $\beta > 0$, differentiating (3) with respect to $\mathbf{x}$ yields:

$$\nabla_{\mathbf{x}} \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}, \Delta) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} \right)^\top E_\beta(\mathbf{x}, \mathbf{y}) \ , \tag{5}$$

where $E_\beta(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\partial \mathbf{dtw}_\beta}{\partial \Delta}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{A \in \mathcal{A}_{T_1,T_2}} e^{-\frac{\langle \mathbf{A}, \Delta(\mathbf{x},\mathbf{y}) \rangle}{\beta}} A}{\sum_{A \in \mathcal{A}_{T_1,T_2}} e^{-\frac{\langle \mathbf{A}, \Delta(\mathbf{x},\mathbf{y}) \rangle}{\beta}}}$ can be interpreted as a weighted average alignment. To compute $E_\beta(\mathbf{x}, \mathbf{y})$, Cuturi and Blondel (2017) proposed to back propagate the forward recursion of Algorithm 1, starting from E_{T_1,T_2} down to $E_{0.0}$. Indeed, the value of $\mathbf{dtw}_\beta$ is stored in the last alignment cost r_{T_1,T_2}. Thus differentiating $\mathbf{dtw}_\beta$ with respect to any $r_{i,j}$ only involves the terms of $r_{i-1,j}, r_{i,j-1}$ and $r_{i-1,j-1}$. Differentiating the softmin operation of the forward pass yields the backward recursion of Algorithm 2.

2.2 Delannoy numbers

Delannoy numbers arise naturally when working with Dynamic time warping: they correspond to the number of feasible alignments $\mathcal{A}_{T_1,T_2}$ and are usually denoted by $D(T_1 - 1, T_2 -$

Algorithm 2 Backward recursion to differentiate $\mathbf{dtw}_\beta$ (Cuturi and Blondel, 2017).

Input: $\mathbf{x}, \mathbf{y}$, parameter β, distance δ and intermediary alignment matrix R
Output: $E = E_\beta(\mathbf{x}, \mathbf{y})$
$r_{i,m+1} = r_{n+1,j} = -\infty,\ i \in [\![n]\!]\,,\ j \in [\![m]\!]$
$e_{i,m+1} = r_{n+1,j} = 0,\ i \in [\![n]\!]\,,\ j \in [\![m]\!]$
$\delta_{i,m+1} = \delta_{n+1,j} = 0,\ i \in [\![n]\!]\,,\ j \in [\![m]\!]$
$\delta_{n+1,m+1} = 0, e_{n+1,m+1} = 1, r_{n+1,m+1} = r_{n,m}$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** m **do**
    $a = \exp \frac{1}{\beta}(r_{i+1,j} - r_{i,j} - \delta_{i+1,j})$
    $b = \exp \frac{1}{\beta}(r_{i,j+1} - r_{i,j} - \delta_{i,j+1})$
    $c = \exp \frac{1}{\beta}(r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1})$
    $e_{i,j} = ae_{i+1,j} + be_{i,j+1} + ce_{i+1,j+1}$
  end for
end for

1) (Cuturi, 2011). The role played by Delannoy numbers is even more apparent when working with Soft-DTW which is only natural since the minimum is replaced with an exponential weighted sum over all feasible alignments. In this section, we provide some new properties of Delannoy numbers which are crucial to establish the time sensitivity of $\mathbf{dtw}_\beta$ and define a heuristic to set the hyperparameter β.

For the sake of convenience, we consider the shifted Delannoy sequence starting at $n = m = 1$ so that: $\mathrm{card}(\mathcal{A}_{m,n}) = D_{m,n}$ for all integers $m, n \geq 1$. Formally, the Delannoy sequence can be defined recursively by:

Definition 3 (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 $\forall m, n \in \mathbb{N}^\star$:*

$$D_{1,n} = D_{m,1} = 1 \tag{6}$$
$$D_{m+1,n+1} = D_{m,n+1} + D_{m+1,n} + D_{m,n}\ . \tag{7}$$

Janati et al. (2020a) showed that there exists a positive constant c such that the diagonal elements $D_{m,m}$ verify the bounded growth inequality $D_{m+1,m+1} \leq c^2 D_{m,m}$. The following proposition provides tighter inequalities that bound the growth of the central Delannoy sequence from both ends. The proof can be found in the supplementary materials.

Proposition 4 *Let $c = 1 + \sqrt{2}$. The central (diagonal) Delannoy sequence $D_m \overset{\text{def}}{=} D_{m,m}$ verifies:*

$$\frac{D_{m+1}}{D_m} \leq c^2 \frac{m}{m + \frac{1}{2}} \quad \forall m \geq 1\ , \tag{8}$$
$$\frac{D_{m+1}}{D_m} \geq c^2 \frac{m}{m + \frac{1}{2} + \frac{1}{2m}} \quad \forall m \geq 1\ . \tag{9}$$

Moreover, these bounds are tight asymptotically since $\frac{D_{m+1}}{D_m} = c^2 \left(+o(\frac{1}{m})\right)$.

Reiterating this inequality over the course of the time series leads to the following corollary which is crucial to derive a simple heuristic to set the $\mathbf{dtw}_\beta$ hyperparameter β in practice. A detailed derivation is provided in the supplementary materials.

Corollary 5 *Let $T > m \geq 1$, $c = 1 + \sqrt{2}$. The central Delannoy numbers verify:*

$$\begin{aligned} c^{2(T-m)} \frac{D_m}{D_T} &\geq \left(\frac{T}{me}\right)^{\frac{1}{2}} \quad \text{for } m \geq 1 \ , \\ c^{2(T-m)} \frac{D_m}{D_T} &\leq e^{\frac{\pi^2}{6} - \frac{5}{8}} \left(\frac{T-1}{m-1}\right)^{\frac{1}{2}} \quad \text{for } m \geq 2 \ . \end{aligned} \tag{10}$$

We conclude this section on Delannoy numbers by restating a result of (Janati et al., 2020a) which will be useful later on. It provides bounds on the off-diagonal Delannoy numbers which lead to the quadratic lower bound of Soft-DTW.

Proposition 6 *Let $c = 1 + \sqrt{2}$. $\forall m, i \in \mathbb{N}^\star$:*

$$D_{m,m+i} \leq c \Phi_{m,i} D_{m,m+i-1} \ , \tag{11}$$

$$c \Psi_{m,i} D_{m,m+i} \leq D_{m+1,m+i} \ , \tag{12}$$

where

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

Moreover, combining the inequalities above with a running $i = 1..k$ we get for any $m, m', k \in \mathbb{N}^\star$ such that $m + m' \leq T - 1$ and $k \leq \min(T - m, m' - 1)$:

$$\log\left(\frac{D_{m,m} D_{m',m'}}{D_{m+k,m} D_{m'-k,m'}}\right) \geq P(k) \ , \tag{13}$$

where $P(k) = \alpha k(k-1) + \rho k + \frac{1}{3T}$ with $\alpha = \frac{2-\sqrt{2}}{T} > 0$ and $\rho = \frac{3\sqrt{2}-4}{3T} > 0$.

2.3 Time sensitivity of Soft-DTW and effect of β

In (Janati et al., 2020a), Soft-DTW was shown to increase quadratically with the size of the temporal shift between two time series. Formally, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^T$ such that $\mathbf{y}$ is temporally shifted w.r.t. $\mathbf{x}$ by k time steps :

$$\mathbf{y}_{i+k} = \mathbf{x}_i \quad \forall i \in [\![1, T-k]\!] \ . \tag{14}$$

Then the following holds:

Proposition 7 *Let $r = \min_{i,j}\{\Delta(\mathbf{x},\mathbf{x})_{ij} | \Delta(\mathbf{x},\mathbf{x})_{ij} > 0\}$. Denote $m = \arg\min_{i \in [\![1,T-1]\!]} \{\mathbf{x}_{i+1} \neq \mathbf{x}_i\}$ $m' = T - \arg\max_{i \in [\![1,T-1]\!]} \{\mathbf{x}_{i+1} \neq \mathbf{x}_i\}$.*

If $0 < \beta \leq \frac{r}{\log(3T D_{T,T})}$:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq \beta \log\left(\frac{D_{m,m} D_{m',m'}}{D_{m+k,m} D_{m'-k,m'}}\right) - \frac{\beta}{3T} \ . \tag{15}$$

PROOF. see (Janati et al., 2020a) for a proof.

An example of a temporal 50-shift is illustrated in Figure 1. The heatmap of the squared Euclidean cost matrix Δ shows two white rectangles 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. Shifting $\mathbf{y}$ temporally would move the red horizontal line downwards, changing the set of alignments with cost 0 without affecting the $\mathbf{dtw}_0$ value (0). 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. This number of equivalent paths is expressed as the product of Delannoy numbers as showed by proposition 7 which, combined with proposition 6 provides a quadratic lower bound for temporal shifts.

A tighter lower bound While the previously considered time series covered a wide range of scenarios, the obtained result requires β to be too small, thereby not providing any insight on how $\mathbf{dtw}_\beta$ behaves when β increases. In the following paragraph, we relax this assumption on β in order to find a tighter lower bound than the one given in (15). We consider the simplified setting of Dirac univariate time series $\mathbf{x}, \mathbf{y}$ such that $\mathbf{y}$ is ahead of $\mathbf{x}$ by k time steps. Formally, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^T$ such that for some $t^\star \in [\![1, T]\!]$ and $1 \leq k \leq T - t^\star$:

$$\begin{aligned} t \neq t^\star &\Rightarrow \mathbf{x}_t = 0 \\ t \neq t^\star + k &\Rightarrow \mathbf{y}_t = 0 \\ \mathbf{x}_{t^\star} &= \mathbf{y}_{t^\star + k} = g \in \mathbb{R} \end{aligned} \tag{16}$$

This simplified setting allows for tighter bounds of Soft-DTW.

Proposition 8 *Consider* $\mathbf{x}$ *and* $\mathbf{y}$ *as defined in* (16). *Let* $r = \delta(g, 0)$, $T \geq 6$, $c = 1 + \sqrt{2}$. *Then:*

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq -\beta \log \left(e^{-P(k)}(1 - \lambda_\beta) + \lambda_\beta H \right) \quad , \tag{17}$$

where: $\lambda_\beta = e^{-\frac{r}{\beta}}$, $H = \frac{c^6 e^{\frac{\pi^2}{6} - \frac{1}{8}}}{6} T^{\frac{1}{2}}$ *and* P *is the quadratic bound defined in Proposition 6.*

PROOF. See the supplementary materials.

Effect of β When β is small enough, λ_β goes to 0, thus the lower bound of Proposition 8 can be very well approximated by the quadratic bound $P(k)$. However when β increases, $\lambda_\beta H$ increases which will dominate the log argument for a sufficiently large k. Using the example of Figure 1 (left), we compute both sides of Equation (17) for 3 values of β. Figure 1 shows that $\mathbf{dtw}_\beta$ saturates for a certain temporal shift $k_{\max}$ beyond which it is no longer sensitive to temporal lags. This phase transition is also observed by the lower bound (17). This provides a heuristic to set β based on a predefined $k_{\max}$ i.e the largest temporal shift the user is willing to capture. Notice that such a point does not always exist (when β is too small) as it may be larger than the time series length T (see top example, Figure 1).

Proposition 9 *Let* $T \geq 3$, $1 \leq k_{\max}$ *and* $0 < \eta < 1$. *Using the same notations of Proposition 8, define the lower bound function:*

$$\mathrm{LB}_\beta : k \mapsto -\beta \log \left(e^{-P(k)}(1 - \lambda_\beta) + \lambda_\beta H \right) \quad .$$

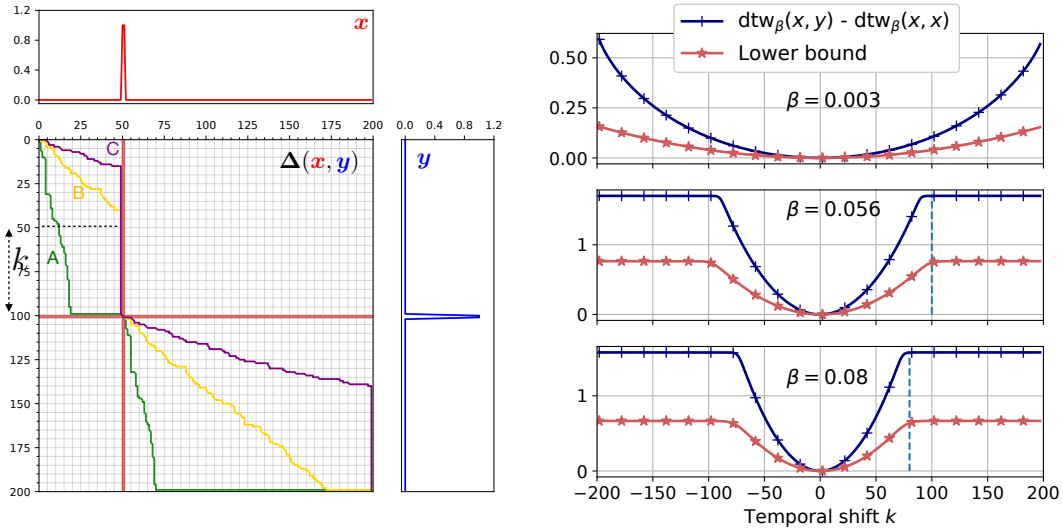

Fig. 1. Left: example of 3 DTW alignment paths (A, B and C) between $\mathbf{x}$ and $\mathbf{y}$ with a temporal 50-shift. The heatmap of the distance matrix Δ (here squared Euclidean) shows 2 red bars where the distance is not equal to 0 except at their intersection. An alignment path has 0 cost if and only if it does not cross the red lines. Compared to the figure of (Janati et al., 2020a), here we consider the simpler case of Dirac time series. This setting allows one to obtain a bound for all values of β, including the asymptotes for large temporal shifts displayed on the right side figure above.

If $\beta \geq \frac{r}{P(k_{\max})+\log((e^{\eta}-1)H)}$ *then:*

$$0 \leq \frac{\lim\limits_{k\to+\infty} \mathrm{LB}_\beta(k) - \mathrm{LB}_\beta(k_{\max})}{\beta} \leq \eta \ . \tag{18}$$

PROOF. See the supplementary materials.

Proposition 9 provides a sufficient condition to set β such that the lower bound LB saturates for a certain $k_{\max}$. In the examples shown in Figure 1, β was set using this heuristic with $\eta = 0.01$ and $k_{\max} \in \{500, 100, 80\}$ respectively top to bottom. The dotted vertical lines highlight the choice of $k_{\max}$ which is very close to the saturation point of $\mathbf{dtw}_\beta$. In practice, we use the same heuristic by setting $r = \max_{i,j} \Delta(\mathbf{x}_i, \mathbf{y}_j)$.

3. Soft-DTW barycenters via alternating optimization

Consider a dataset with N multivariate time series $\mathbf{x}_1, \ldots, \mathbf{x}_N$ assumed to have the same dimension p and respective time lengths $T_1, \ldots, T_N$. Let $w_1, \ldots, w_n$ be a set of positive weights summing to one. The Soft-DTW barycenter with cost Δ and fixed length T is defined

as:

$$\bar{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^{p,T}}{\arg\min} \sum_{i=1}^{N} w_i \mathbf{dtw}_\beta(\mathbf{x}_i, \mathbf{x}, \Delta) \tag{19}$$

$$= \underset{\mathbf{x}\in\mathbb{R}^{p,T}}{\arg\min} -\sum_{i=1}^{N} w_i \beta \log \left(\sum_{A\in\mathcal{A}_{T,T_i}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{x}_i)\rangle}{\beta}} \right) \ . \tag{20}$$

Alternating optimization Provided that Δ is differentiable, the most straightforward solution to (19) would probably be to use a Quasi-Newton method. However, since we intend to use an OT loss for Δ, computing each gradient step would require $T\sum_{i=1}^{N} T_i$ OT gradients. Instead, we use Fenchel duality to obtain an alternating optimization problem that not only avoids the computation of the gradients of Δ but also spares us any form of step-size backtracking. This is given by proposition 10.

Proposition 10 *For the sake of convenience, provide the sets of binary matrices $\mathcal{A}_{T_i,T}$ with some arbitrary indexation $\mathcal{A}_{T_i,T} = \{A_i^1, \dots, A_i^{D_{T_i,T}}\}$ and let S_K denote the probability simplex of $\mathbb{R}^K$. The Soft-DTW problem (19) is equivalent to the optimization problem:*

$$\min_{\mathbf{x}\in\mathbb{R}^{p,T}} \min_{\substack{\theta_1\in S_{D_{T_1,T}} \\ \dots \\ \theta_N\in S_{D_{T_N,T}}}} \sum_{i=1}^{N} w_i \left[\langle \sum_{k=1}^{D_{T_k,T}} \theta_i^k A_i^k, \Delta(\mathbf{x}_i, \mathbf{x})\rangle + \beta H(\theta_i) \right] \ . \tag{21}$$

PROOF. A standard result in convex optimization theory states that the Fenchel conjugate of entropy is logsumexp. Formally, for $\mathbf{x} \in \mathbb{R}^K$:

$$(\beta H)^\star(\mathbf{x}) \overset{\text{def}}{=} \max_{\theta\in S_K} \langle \mathbf{x}, \theta\rangle - \beta H(\theta) = \beta \log \left(\sum_{k=1}^{K} e^{-\frac{\mathbf{x}_k}{\beta}} \right) \ . \tag{22}$$

Thus:

$$-\beta \log \left(\sum_{k=1}^{K} e^{-\frac{\mathbf{x}_k}{\beta}} \right) = \min_{\theta\in S_K} \langle -\mathbf{x}, \theta\rangle + \beta H(\theta) \ . \tag{23}$$

Therefore, the barycenter loss (19) can be written:

$$\min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} w_i \mathbf{dtw}_\beta(\mathbf{x}_i, \mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} -w_i \beta \log \left(\sum_{A\in\mathcal{A}_{T,T_i}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{x}_i)\rangle}{\beta}} \right) \tag{24}$$

$$= \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} w_i \min_{\theta_i\in S_{D_{T_i,T}}} \langle \theta_i, \left(\langle A_i^1, \Delta(\mathbf{x}_i, \mathbf{x})\rangle, \dots, \langle A^{D_{T_i,T}}{}_i, \Delta(\mathbf{x}_i, \mathbf{x})\rangle\right)\rangle + \beta H(\theta_i) \tag{25}$$

$$= \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} w_i \min_{\theta_i\in S_{D_{T_i,T}}} \sum_{k=1}^{D_{T_k,T}} \langle \theta_i^k A_i^k, \Delta(\mathbf{x}_i, \mathbf{x})\rangle + \beta H(\theta_i) \ , \tag{26}$$

where the last equality follows from the separability of the sum with respective to the θ_i ■

The major benefit of the dual formulation of Proposition (10) is the ability to compute Fréchet means of Δ directly. This will be in particular crucial when we define Δ as an OT divergence for which Fréchet means are orders of magnitude faster to compute using Sinkhorn's algorithm than via gradient based methods (Cuturi and Peyré, 2018). While minimizing with respect to the θ_i seems computationally unfeasible due their large dimension, their update is actually not required to compute the new $\mathbf{x}$. Instead, one needs to update the matrices $\mathbf{Z}_i \stackrel{\text{def}}{=} \sum_{k=1}^{D_{T_k,T}} \theta_i^k A_i^k$ which are *exactly* given by the gradients $\frac{\partial \mathbf{dtw}_\beta(\mathbf{x}_i,\mathbf{x})}{\partial \Delta}(\mathbf{x}_i,\mathbf{x})$. Indeed, given that the loss is convex in θ_i, for a fixed $\mathbf{x}$, the optimal θ_i verifies the KKT conditions for some Lagrange multiplier λ_i:

$$\begin{cases} \langle A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle + \beta \log(\theta_i^k) - \lambda_i = 0 \\ \sum_{k=1}^{D_{T_i,T}} \theta_i^k = 1 \ , \end{cases}$$

which leads to:

$$\theta_i^k = \frac{e^{-\frac{\langle A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}}}{\sum_{k=1}^{D_{T_k,T}} e^{-\frac{\langle A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}}} \ . \tag{27}$$

Thus:

$$\mathbf{Z}^i \stackrel{\text{def}}{=} \sum_{k=1}^{D_{T_k,T}} \stackrel{\text{def}}{=} \theta_i^k A_i^k = \frac{\sum_{k=1} e^{-\frac{\langle A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}} A_i^k}{\sum_{k=1}^{D_{T_k,T}} e^{-\frac{\langle A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}}} = \frac{\partial \mathbf{dtw}_\beta(\mathbf{x}_i,\mathbf{x})}{\partial \Delta}(\mathbf{x}_i,\mathbf{x}) \ , \tag{28}$$

which can be computed using Algorithm 2. Notice that to update $\mathbf{x}$, computing θ_i is not necessary: it is sufficient to have updated $\mathbf{Z}_i$. This leads to algorithm 3.

Initialization It is important to keep in mind the loss (21) is not jointly convex in $\mathbf{x}$ and θ. Thus, algorithm 3 is not guaranteed to converge to a global minimum. Nevertheless, in our experiments, initializing $\mathbf{x}$ with a uniform distribution leads to meaningful barycenters with the desired spatio-temporal properties.

Algorithm 3 Soft-DTW barycenter.

Input: $\mathbf{x}_1, \ldots, \mathbf{x}_N$, $\mathbf{x}_0$, weights $w_1, \ldots, w_N$, parameter β
Output: solution of (21)
Initialize $\mathbf{x} = \mathbf{x}_0 \in \mathbb{R}^{p,T}$, **compute** $\Delta(\mathbf{x}_i, \mathbf{x})$ **for all** $i = 1..N$
while not converged **do**
  for $i = 1$ **to** N **do**
    Compute $\mathbf{Z}^i$ with Algorithm 2
  end for
  for $t = 1$ **to** T **do**
    $\mathbf{x}^t = \arg\min_{\mathbf{a}\in\mathbb{R}^p} \sum_{i=1}^N \sum_{t'=1}^{T_i} w_i \mathbf{Z}^i_{t',t} \Delta(\mathbf{x}_i^{t'}, \mathbf{a})$
  end for
end while

4. Spatio-temporal barycenters

4.1 $\widetilde{\text{UOT}}$ as debiased unbalanced OT

We start this section by explaining how we use optimal transport to define the alignment cost $\widetilde{\text{UOT}}$ between two non-negative measures $\mathbf{x}, \mathbf{y}$ with a fixed support given by $\mathcal{A} = \{a_1, \dots, a_p\} \subset \mathbb{R}^d$. Since the support is fixed, $\mathbf{x}, \mathbf{y}$ can be identified with vectors of non-negative weights i.e $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^p$.

Entropy regularized unbalanced OT and Sinkhorn's algorithm Let $\mathbf{C} \in \mathbb{R}_+^{d,d}$ be the pairwise distance matrix given by $\mathbf{C}_{ij} = c(a_i, a_j)$, where c is a symmetric Lipschitz cost function such that $\exp(-\frac{\mathbf{C}}{\varepsilon})$ is positive semi-definite for any $\varepsilon > 0$. The matrix $\mathbf{C}$ – known as the ground metric – defines the geometry that OT distances lift to compute transportation costs. Formally, transporting a fraction of mass $\mathbf{P}_{ij}$ from a_i to a_j is given by $\mathbf{P}_{ij}\mathbf{C}_{ij}$, the total cost of transport is given by $\langle \mathbf{P}, \mathbf{C} \rangle = \sum_{ij} \mathbf{P}_{ij}\mathbf{C}_{ij}$. To guarantee mass transportation, Liero et al. (2016) introduced the following formulation of unbalanced OT:

$$\min_{\mathbf{P} \in {\mathbb{R}_+}^{p \times p}} \langle \mathbf{P}, \mathbf{C} \rangle + \gamma \text{KL}(\mathbf{P}\mathbb{1}|\mathbf{x}) + \gamma \text{KL}(\mathbf{P}^\top \mathbb{1}|\mathbf{y}) \ , \tag{29}$$

where $\gamma > 0$ is a fixed hyperparameter. Chizat et al. (2017) generalized the above problem to other divergences than KL while adding entropy regularization (Cuturi, 2013). Up to an additional constant, this generalized problem is equivalent to:

$$\text{UOT}(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{P} \in {\mathbb{R}_+}^{p \times p}} \varepsilon \text{KL}(\mathbf{P}|e^{-\frac{\mathbf{C}}{\varepsilon}}) + \gamma \text{KL}(\mathbf{P}\mathbb{1}|\mathbf{x}) + \gamma \text{KL}(\mathbf{P}^\top \mathbb{1}|\mathbf{y}) \ . \tag{30}$$

The following proposition – taken from (Janati et al., 2020a) – shows how to compute UOT using a generalized version of Sinkhorn's algorithm.

Proposition 11 *(Janati et al., 2020a) Problem (30) is equivalent to:*

$$\text{UOT}(\mathbf{x}, \mathbf{y}) = \max_{u,v \in \mathbb{R}^p} -\gamma \langle \mathbf{x}, e^{-\frac{u}{\gamma}} - 1 \rangle - \gamma \langle \mathbf{y}, e^{-\frac{v}{\gamma}} - 1 \rangle - \varepsilon \langle e^{\frac{u \oplus v}{\varepsilon}} - 1, e^{-\frac{\mathbf{C}}{\varepsilon}} \rangle \ . \tag{31}$$

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

$$\mathbf{a} = \left(\frac{\mathbf{x}}{\mathbf{K}\mathbf{b}}\right)^\omega \quad , \quad \mathbf{b} = \left(\frac{\mathbf{y}}{\mathbf{K}^\top \mathbf{a}}\right)^\omega \tag{32}$$

In particular, if $\mathbf{x} = \mathbf{y}$, *then there exists a pair of solutions such that* $\mathbf{a} = \mathbf{b}$.

Solving the fixed point problem (32) is equivalent to alternate maximization of the dual function (31). Starting from two vectors $\mathbf{a}, \mathbf{b}$ set to $\mathbb{1}$, the algorithm iterates through the scaling operations (32). This is a generalization of the Sinkhorn algorithm which corresponds to $\omega = 1$ or $\gamma = +\infty$.

PROOF. The symmetry of the dual problem (31) with $\mathbf{x} = \mathbf{y}$ implies immediately that $\mathbf{a} = \mathbf{b}$. Proposition 11 gives the fixed point equation.

Debiased UOT Due to regularization, UOT is known to suffer from an entropy bias causing UOT to fail to identify identical distributions $\mathrm{UOT}(\mathbf{x}, \mathbf{x}) \neq 0$. Inspired by the balanced OT case Feydy et al. (2018), we define the debiased UOT cost:

$$\widetilde{\mathrm{UOT}}(\mathbf{x}, \mathbf{y}) = \mathrm{UOT}(\mathbf{x}, \mathbf{y}) - \frac{1}{2}\left(\mathrm{UOT}(\mathbf{x}, \mathbf{x}) + \mathrm{UOT}(\mathbf{y}, \mathbf{y})\right) \quad . \tag{33}$$

$\widetilde{\mathrm{UOT}}$ is coercive with respect to each of its arguments, moreover, if $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$ is positive semi-definite, then $\widetilde{\mathrm{UOT}}$ is non-negative (Janati et al., 2020a). Therefore, $\mathbf{dtw}_\beta$ is well defined with $\widetilde{\mathrm{UOT}}$ as a cost function, a loss function we named STA: *Spatio-Temporal Alignment*.

Definition 12 (STA) *We define the STA loss as:*

$$\mathbf{sta}_\beta(\mathbf{x}, \mathbf{y}) = \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}; \widetilde{\mathrm{UOT}}) \quad . \tag{34}$$

Remark 13 *It is important to note that all the properties shown for $\widetilde{\mathrm{UOT}}$ (coercivity, non-negativity) hold only if the reference measure used to define the entropy regularization is the uniform distribution over a fixed universal support $\mathcal{A}$ for all measures i.e the penalty applied to the transportation plan writes $E(\mathbf{P}) \overset{\text{def}}{=} KL(\mathbf{P}|\mathbb{1}_p/p)$. For instance, when considering the more general formulation with the product measure as a reference ($KL(\mathbf{P}|\mathbf{x} \otimes \mathbf{y})$), Séjourné et al. (2019) showed that non-negativity and convexity hold but with an additional quadratic term between the masses of the distributions $\frac{\varepsilon}{2}(\mathbf{x}^\top \mathbb{1} - \mathbf{y}^\top \mathbb{1})^2$ in the definition of $\widetilde{\mathrm{UOT}}$.*

4.2 Computing barycenters with $\widetilde{\mathrm{UOT}}$

To use algorithm 3, it is required to compute barycenters with the inner $\mathbf{dtw}_\beta$ cost function $\widetilde{\mathrm{UOT}}$ as dissimilarity. In the following, we show that this barycenter can be estimated using a modified Sinkhorn algorithm. Let $\mathbf{x}_1, \ldots, \mathbf{x}_K \in \mathbb{R}_+^p$ and $w_1, \ldots, w_K$ a sequence of positive weights adding to 1. $\widetilde{\mathrm{UOT}}$ is non-negative and coercive, thus its barycenter problem is well defined:

$$\min_{\mathbf{x} \in \mathbb{R}_+^p} \mathcal{J}(\mathbf{x}) \overset{\text{def}}{=} \min_{\mathbf{x} \in \mathbb{R}_+^p} \sum_{k=1}^K w_k \widetilde{\mathrm{UOT}}(\mathbf{x}_k, \mathbf{x}) \quad . \tag{35}$$

$\widetilde{\mathrm{UOT}}$ is differentiable, and its gradient is given by $\gamma(1 - \mathbf{a}^{-\frac{\varepsilon}{\gamma}}, 1 - \mathbf{b}^{-\frac{\varepsilon}{\gamma}})$ where $(\mathbf{a}, \mathbf{b})$ is the solution of the fixed point equation (32) Feydy et al. (2017). Thus, using the chain rule, $\mathcal{J}$ is also differentiable and its gradient is given by:

$$\nabla \mathcal{J}(\mathbf{x}) = \gamma(\mathbf{c}^{-\frac{\varepsilon}{\gamma}} - \sum_{k=1}^K w_k \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}}) \quad , \tag{36}$$

where, using the notations of Proposition 11, $\mathbf{c}, \mathbf{b}_1, \ldots, \mathbf{b}_K, \mathbf{a}_1, \ldots, \mathbf{a}_K \in \mathbb{R}_+^p$ verify the fixed point equations:

$$\mathbf{a}_k = \left(\frac{\mathbf{x}_k}{\mathbf{K}\mathbf{b}_k}\right)^\omega \quad , \quad \mathbf{b}_k = \left(\frac{\mathbf{x}}{\mathbf{K}^\top \mathbf{a}_k}\right)^\omega \quad , \quad \mathbf{c} = \left(\frac{\mathbf{x}}{\mathbf{K}\mathbf{c}}\right)^\omega \tag{37}$$

Without studying the convexity of $\mathcal{J}$, we can show that any stationary point of $\mathcal{J}$ is actually a global minimum. Thus, it is sufficient to solve $\nabla\mathcal{J}(\mathbf{x}) = 0$ to compute the $\widetilde{\text{UOT}}-$barycenter. The following lemma plays a major role in proving this statement.

Lemma 14 (Suboptimality) *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^p$. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^p$ be the optimal dual variables associated with $\text{UOT}(\mathbf{x}, \mathbf{y})$ i.e the solutions of the optimality conditions $\mathbf{a} = \left(\frac{\mathbf{x}}{\mathbf{Kb}}\right)^\omega$ and $\mathbf{b} = \left(\frac{\mathbf{y}}{\mathbf{K}^\top\mathbf{a}}\right)^\omega$. Then for any $\mathbf{f}, \mathbf{g} \in \mathbb{R}_+^p$:*

$$\gamma\langle \mathbf{x}, \mathbf{f}^{-\frac{\varepsilon}{\gamma}}\rangle + \gamma\langle \mathbf{y}, \mathbf{g}^{-\frac{\varepsilon}{\gamma}}\rangle + \varepsilon\langle \mathbf{f}, \mathbf{Kg}\rangle \geq (\varepsilon + 2\gamma)\langle \mathbf{a}, \mathbf{Kb}\rangle \ . \tag{38}$$

PROOF. Using the same change of variable $\mathbf{f} = e^{\frac{u}{\varepsilon}}, \mathbf{g} = e^{\frac{v}{\varepsilon}}$, the dual problem of Proposition 11 can be written:

$$\begin{aligned}
\text{UOT}(\mathbf{x}, \mathbf{y}) &= \max_{\mathbf{f},\mathbf{g}\in\mathbb{R}_+^p} -\gamma\langle \mathbf{x}, \mathbf{f}^{-\frac{\varepsilon}{\gamma}} - 1\rangle - \gamma\langle \mathbf{y}, \mathbf{g}^{-\frac{\varepsilon}{\gamma}} - 1\rangle - \varepsilon\langle \mathbf{f}\otimes\mathbf{g} - 1, \mathbf{K}\rangle \\
&= \max_{\mathbf{f},\mathbf{g}\in\mathbb{R}_+^p} -\gamma\langle \mathbf{x}, \mathbf{f}^{-\frac{\varepsilon}{\gamma}}\rangle - \gamma\langle \mathbf{y}, \mathbf{g}^{-\frac{\varepsilon}{\gamma}}\rangle - \varepsilon\langle \mathbf{f}, \mathbf{Kg}\rangle - \varepsilon\|\mathbf{K}\|_1 + \gamma(\|\mathbf{x}\|_1 + \|\mathbf{y}\|_1) \ .
\end{aligned}$$

Since $\mathbf{a}, \mathbf{b}$ are the solution of the dual problem above, at optimality it holds:

$$\langle \mathbf{x}, \mathbf{a}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{a}^{\frac{1}{\omega}}\mathbf{Kb}, \mathbf{a}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{a}, \mathbf{Kb}\rangle \ .$$

Similarly:

$$\langle \mathbf{y}, \mathbf{b}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{b}^{\frac{1}{\omega}}\mathbf{K}^\top\mathbf{a}, \mathbf{b}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{b}, \mathbf{K}^\top\mathbf{a}\rangle = \langle \mathbf{a}, \mathbf{Kb}\rangle \ .$$

Thus:

$$\text{UOT}(\mathbf{x}, \mathbf{y}) = -(\varepsilon + 2\gamma)\langle \mathbf{a}, \mathbf{Kb}\rangle - \varepsilon\|\mathbf{K}\|_1 + \gamma(\|\mathbf{x}\|_1 + \|\mathbf{y}\|_1) \ .$$

By the definition of the max operator, it holds for any $\mathbf{f}, \mathbf{g} \in \mathbb{R}_+^p$:

$$\gamma\langle \mathbf{x}, \mathbf{f}^{-\frac{\varepsilon}{\gamma}}\rangle + \gamma\langle \mathbf{y}, \mathbf{g}^{-\frac{\varepsilon}{\gamma}}\rangle + \varepsilon\langle \mathbf{f}, \mathbf{Kg}\rangle \geq (\varepsilon + 2\gamma)\langle \mathbf{a}, \mathbf{Kb}\rangle \ .$$

□

Since $\mathcal{J}$ is coercive, it has at least one global minimum. The following proposition shows that this minimum is unique.

Proposition 15 *Let $\bar{\mathbf{x}} \in \mathbb{R}_+^p$ such that $\nabla\mathcal{J}(\bar{\mathbf{x}}) = \mathbf{0}$. Then for any $\mathbf{y} \in \mathbb{R}_+^p$ it holds:*

$$\mathcal{J}(\mathbf{y}) \geq \mathcal{J}(\bar{\mathbf{x}}).$$

PROOF. Let $\mathbf{d}_1, \ldots, \mathbf{d}_K$ be the symmetric dual variables used to compute $\widetilde{\text{UOT}}(\mathbf{x}_k, \mathbf{x}_k)$ for $k = 1..K$ i.e the solutions of $\mathbf{d}_k = \left(\frac{\mathbf{x}_k}{\mathbf{Kd}_k}\right)^\omega$. Let $\mathbf{y} \in \mathbb{R}_+^p$ and its associated dual variables $\mathbf{c}', \mathbf{a}'_1, \ldots, \mathbf{a}'_K, \mathbf{b}'_1, \ldots, \mathbf{b}'_K$ used to compute $\widetilde{\text{UOT}}(\mathbf{y}, \mathbf{y}), \widetilde{\text{UOT}}(\mathbf{x}_1, \mathbf{y}), \ldots, \widetilde{\text{UOT}}(\mathbf{x}_K, \mathbf{y})$. Therefore, it holds:

$$\mathcal{J}(\mathbf{y}) = (\varepsilon + 2\gamma)\sum_{k=1}^K w_k \left(\frac{1}{2}(\langle \mathbf{c}', \mathbf{Kc}'\rangle + \langle \mathbf{d}_k, \mathbf{Kd}_k\rangle) - \langle \mathbf{a}'_k, \mathbf{Kb}'_k\rangle\right) \ .$$

Let $\mathbf{c}, \mathbf{a}_1, \dots, \mathbf{a}_K, \mathbf{b}_1, \dots, \mathbf{b}_K$ denote the dual variables verifying (37) for $\mathbf{x} = \bar{\mathbf{x}}$. Moreover, $\nabla \mathcal{J}(\bar{\mathbf{x}}) = \mathbf{0}$ leads to $\mathbf{c}^{-\frac{\varepsilon}{\gamma}} = \sum_{k=1}^K w_k \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}}$. Therefore:

$$\sum_{k=1}^K w_k \langle \mathbf{a}_k, \mathbf{K}\mathbf{b}_k \rangle = \sum_{k=1}^K w_k \langle \bar{\mathbf{x}}, \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \bar{\mathbf{x}}, \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \ . \tag{39}$$

Thus, evaluate $\mathcal{J}$ at $\bar{\mathbf{x}}$ leads to:

$$\begin{aligned}\mathcal{J}(\bar{\mathbf{x}}) &= (\varepsilon + 2\gamma) \sum_{k=1}^K w_k \left(\frac{1}{2}(\langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle + \langle \mathbf{d}_k, \mathbf{K}\mathbf{d}_k \rangle) - \langle \mathbf{a}_k, \mathbf{K}\mathbf{b}_k \rangle \right) \\ &= \frac{1}{2}(\varepsilon + 2\gamma) \left(\sum_{k=1}^K w_k \langle \mathbf{d}_k, \mathbf{K}\mathbf{d}_k \rangle - \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \right) \ .\end{aligned}$$

Thus, the statement we wish to prove is equivalent to:

$$\mathcal{J}(\mathbf{y}) \geq \mathcal{J}(\bar{\mathbf{x}}) \Leftrightarrow \frac{1}{2}(\varepsilon + 2\gamma) \left(\langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle + \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \right) \geq (\varepsilon + 2\gamma) \sum_{k=1}^K w_k \langle \mathbf{a}'_k, \mathbf{K}\mathbf{b}'_k \rangle \ . \tag{40}$$

For each element of the sum in the right side above, let us derive an upper bound using Lemma 14. Consider the sub-optimal dual variables $(\mathbf{f}_k, \mathbf{g}_k) = (\mathbf{a}_k, \mathbf{b}_k \odot \frac{\mathbf{c}'}{\mathbf{c}})$. It holds:

$$\gamma \langle \mathbf{x}_k, \mathbf{a}_k^{-\frac{\varepsilon}{\gamma}} \rangle + \gamma \langle \mathbf{y}, (\mathbf{b}_k \odot \frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}} \rangle + \varepsilon \langle \mathbf{b}_k \odot \frac{\mathbf{c}'}{\mathbf{c}}, \mathbf{K}^\top \mathbf{a}_k \rangle \geq (\varepsilon + 2\gamma) \langle \mathbf{a}'_k, \mathbf{K}\mathbf{b}'_k \rangle \ . \tag{41}$$

Applying the weighted sum and using the optimality conditions along with $\mathcal{J}(\bar{\mathbf{x}}) = 0$, the elements in the left side can be further simplified as:

$$\begin{aligned}&\sum_{k=1}^K w_k \langle \mathbf{x}_k, \mathbf{a}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \sum_{k=1}^K w_k \langle \bar{\mathbf{x}}, \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \bar{\mathbf{x}}, \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \\ &\sum_{k=1}^K w_k \langle \mathbf{y}, (\mathbf{b}_k \odot \frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{y} \odot (\frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}}, \sum_{k=1}^K w_k \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{y} \odot (\frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}}, \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{y}, \mathbf{c}'^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle \\ &\sum_{k=1}^K w_k \langle \mathbf{b}_k \odot \frac{\mathbf{c}'}{\mathbf{c}}, \mathbf{K}^\top \mathbf{a}_k \rangle = \langle \frac{\mathbf{c}'}{\mathbf{c}} \odot \bar{\mathbf{x}}, \sum_{k=1}^K w_k \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}' \odot \bar{\mathbf{x}}, \mathbf{c}^{-\frac{\gamma+\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}', \mathbf{K}\mathbf{c} \rangle \ .\end{aligned}$$

Therefore, summing over equation (41):

$$\gamma \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle + \gamma \langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle + \varepsilon \langle \mathbf{c}', \mathbf{K}\mathbf{c} \rangle \geq (\varepsilon + 2\gamma) \sum_{k=1}^K w_k \langle \mathbf{a}'_k, \mathbf{K}\mathbf{b}'_k \rangle \ .$$

On another side, since $\mathbf{K}$ is positive semi-definite, it holds:

$$\langle \mathbf{c} - \mathbf{c}', \mathbf{K}(\mathbf{c} - \mathbf{c}') \rangle \geq 0 \Rightarrow \frac{1}{2}(\langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle + \langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle) \geq \langle \mathbf{c}', \mathbf{K}\mathbf{c} \rangle \ .$$

Combining the last two inequalities leads to (40) ending the proof. □.

To solve the barycenter problem (35), it is sufficient to solve the fixed point system:

$$\mathbf{a}_k = \left(\frac{\mathbf{x}_k}{\mathbf{K}\mathbf{b}_k}\right)^\omega \quad , \quad \mathbf{b}_k = \left(\frac{\bar{\mathbf{x}}}{\mathbf{K}^\top \mathbf{a}_k}\right)^\omega \quad , \quad \mathbf{c} = \left(\frac{\bar{\mathbf{x}}}{\mathbf{K}\mathbf{c}}\right)^\omega \quad , \quad \sum_{k=1}^K w_k \mathbf{b}_k^{-\frac{\varepsilon}{\gamma}} = \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \tag{42}$$

which – combining the last 3 equations – is equivalent to:

$$\mathbf{a}_k = \left(\frac{\mathbf{x}_k}{\mathbf{K}\mathbf{b}_k}\right)^\omega, \quad \mathbf{b}_k = \left(\frac{\bar{\mathbf{x}}}{\mathbf{K}^\top \mathbf{a}_k}\right)^\omega, \quad \mathbf{c} = \left(\frac{\bar{\mathbf{x}}}{\mathbf{K}\mathbf{c}}\right)^\omega, \quad \bar{\mathbf{x}} = \mathbf{c}^{\frac{1}{\omega}} \left(\sum_{k=1}^K w_k (\mathbf{K}^\top \mathbf{a}_k)^{1-\omega}\right)^{\frac{1}{1-\omega}} \tag{43}$$

These equations are very similar to the barycentric Sinkhorn algorithm of Chizat et al. (2017). Indeed, disregarding the symmetric equation in $\mathbf{c}$ and setting $\mathbf{c} = \mathbb{1}_p$ in the update of $\bar{\mathbf{x}}$, we recover Sinkhorn's iterations for the UOT barycenter. These updates lead to Algorithm 4. While the theoretical analysis of its convergence is left for future work, we empirically observe that it converges regardless of the initialization of the dual variables. More importantly, it leads sharper barycenter than the (biased) UOT barycenters for almost no additional computational cost. Figure 2 shows the example of Gaussians with different means, variances and masses for 3 values of ε.

Algorithm 4 Debiased unbalanced $\widetilde{\text{UOT}}$ barycenter.

Input: $\mathbf{x}_1, \dots, \mathbf{x}_K \in \mathbb{R}_+^p$, **parameters** $\varepsilon, \gamma > 0$, $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{M}}{\varepsilon}}$
Output: $\bar{\mathbf{x}}$, **the** $\widetilde{\text{UOT}}$ **barycenter of** $(\mathbf{x}_1, \dots, \mathbf{x}_K)$
Initialize $\mathbf{c} = \mathbf{b}_1 = \cdots = \mathbf{b}_K = \mathbb{1}_p$, **set** $\omega = \frac{\gamma}{\gamma+\varepsilon}$
while Not converged **do**
  for $k = 1$ **to** K **do**
    $\mathbf{a}_k = \left(\frac{\mathbf{x}_k}{\mathbf{K}\mathbf{b}_k}\right)^\omega$
  end for
  $\bar{\mathbf{x}} = \mathbf{c}^{\frac{1}{\omega}} \left(\sum_{k=1}^K w_k (\mathbf{K}^\top \mathbf{a}_k)^{1-\omega}\right)^{\frac{1}{1-\omega}}$
  for $k = 1$ **to** K **do**
    $\mathbf{b}_k = \left(\frac{\bar{\mathbf{x}}}{\mathbf{K}^\top \mathbf{a}_k}\right)^\omega$
  end for
  $\mathbf{c} = \left(\frac{\bar{\mathbf{x}}}{\mathbf{K}\mathbf{c}}\right)^\omega$
end while

Remark 16 *The proposition 15 may seem indicate that $\mathcal{J}$ has a positive curvature. However, it is easy to show that $\widetilde{\text{UOT}}$ is not convex in dimension 1. Indeed, taking $p = 1$ leads to $\mathbf{K} = 1$ and the Sinkhorn equations can be solved in closed form. We obtain for $x, y \in \mathbb{R}_+$:*

$$\widetilde{\text{UOT}}(x, y) = (\varepsilon + 2\gamma)\left(\frac{x^{\frac{2\omega}{\omega+1}} + y^{\frac{2\omega}{\omega+1}}}{2} - (xy)^{\frac{\omega}{\omega+1}}\right).$$

Since $\omega < 1$, in dimension 1, $x \mapsto \widetilde{\text{UOT}}(x, y)$ is strictly concave provided x is large enough.

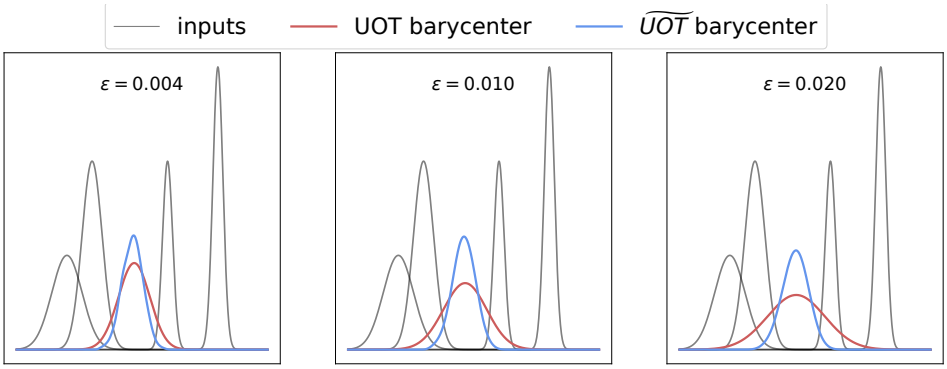

Fig. 2. Barycenters of 4 Gaussian distributions with different means, variances and total masses. The debiased barycenter $\widetilde{\text{UOT}}$ is less sensitive to entropy regularization (ε) than the original unbalanced barycenter UOT.

Complexity analysis As shown by Algorithm 1, soft-DTW is quadratic in time. Computing the $\widetilde{\text{UOT}}$ pairwise 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 Algorithm 4 on 2D data from $O(p^2)$ to $O(p^{\frac{3}{2}})$. Moreover, to leverage fast matrix products on GPUs, computing the $\widetilde{\text{UOT}}$ barycenter and the pairwise $\widetilde{\text{UOT}}$ distance matrices can be done in parallel so that the iterations within both k-loops in Algorithm 4 are run simultaneously.

5. Experiments

We illustrate the effectiveness of the STA barycenter in two experiments with real data by comparing it to the following benchmark methods:

- Euclidean mean;
- DTW barycenter averaging (DBA): the Fréchet mean using DTW as loss function introduced by Petitjean et al. (2011). We use the code provided by the authors https://github.com/fpetitjean/DBA;
- UOT averaging: averaging using OT but disregarding time. Useful to evaluate the importance of the temporal dimension.

Python code for all the experiments can be found in https://github.com/hichamjanati/spatio-temporal-alignements.

Optimal transport hyperparameters The debiased divergence $\widetilde{\text{UOT}}$ is defined by the same hyperparameters of UOT: ε and γ. Given that some of the entropy bias is removed with $\widetilde{\text{UOT}}$, the obtained barycenter is less sensitive to ε than the original UOT barycenter (Fig 2). However, setting ε too large slows down the convergence of the symmetric potential $\mathbf{c}$. Here we set $\varepsilon = 1./p$. 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 that are concentrated far from each other. We set γ at the largest value guaranteeing a minimal transported mass using the heuristic proposed in Janati et al. (2019).

5.1 Averaging of brain imaging data

Studying the function of the various regions of the Human brain is one of the primary goals of neuroimaging research. These studies usually involve a group of healthy individuals (subjects) or patients who perform a series of tasks while having their neural activity recorded from which active regions of the brain are localized. However, drawing conclusions at a population level requires an aggregation function that combines the individual active sources of each subject. While averaging may seem like a straightforward and simple solution, it does not take into account the anatomical differences across subjects which lead to spatially blurred means. Moreover, the brain responses of the different subjects are never synced in time, specially when working with Electro-encephalogrphy (EEG) or Magneto-encephalography (MEG) data which have a high temporal resolution of the order of 1 millisecond. We use the public EEG/MEG dataset DS117 (Wakeman and Henson, 2015) and compute the spatio-temporal source configuration of 6 subjects who were shown images of Human faces using MNE-Python (Gramfort et al., 2013). Here the support of our measures $\mathcal{A}$ is taken to be the set of 642 vertices that define the cortical mesh of the brain. The OT ground metric $\mathbf{C}$ is defined as the quadratic length of the shortest path on the triangulated mesh. We compute 3 different averages: a Euclidean mean, a $\widetilde{\text{UOT}}$ barycenter (independently across time) and a spatio-temporal STA barycenter with $k_{\max} = 20$. As shown in Figure 3, the first burst in the neural response is a visually evoked potential (known as P1) that arises around 100ms after the stimulus (Slotnick et al., 1999) in the primary visual cortex (blue). Then, at around 170ms, an evoked response that is specific to the display of faces occurs in a small region known as the Fusiform Face Area (Green) (FFA) (Bentin et al., 1996; Kanwisher et al., 1997). The delimited regions of interest were selected using the meta-analysis tool Neurosynth (Yarkoni, 2014). As expected, OT based averages (UOT and STA) show less spurious activation patterns but also a smoother temporal transition from P1 to the FFA response than euclidean based ones (Euclidean mean, DBA). To further assess the temporal sensitivity of STA, we display in Figure 4 the ℓ_2 norm across space of the 3 barycenters of Figure 3. The temporal distinction between the two evoked responses is more pronounced with STA than the other methods.

5.2 Forecasting the motion of handwritten letters

Dataset We evaluate the performance of STA in a prediction task using 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 13 time points of (30×30) images for each time series. Each image can thus be seen as a screenshot at a certain time during the writing motion. To make the task a bit more challenging, we randomly shift each time series spatially (resp. temporally) by 0 to 10 pixels in each direction (by keeping 5 to 13 time points evenly selected). The dataset is composed of 20 samples of each one of the letters (“a”, “b”, “c”, “v”), thus the full shape of the dataset is (100, 13, 30, 30).

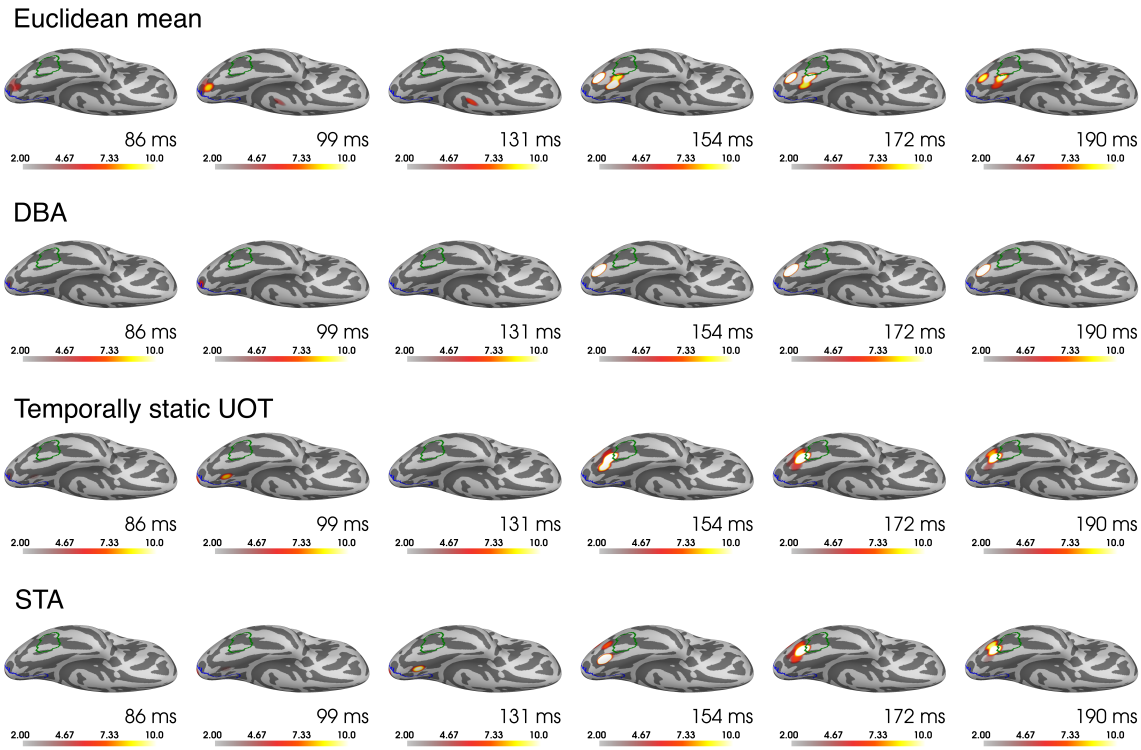

Fig. 3. Barycenters of the spatio-temporal neural activity of 6 subjects taken from the DS117 dataset. The STA barycenter shows a more focused activation around the Fusiform Form Face Area (green) than the other methods. Unlike the OT barycenter, STA shows a more plausible time occurrence of the first evoked response around 100ms.

Forecasting We propose to use barycenters as a forecasting method. For each time series $\mathbf{x}$ in the dataset, knowing only the first $t_0 < 13$ time points, we would like to predict the rest. First, based only on the observed $t_0 = 5$ time points, we select the closest 5 neighbors of $\mathbf{x}$ in the data based on some loss function d. We denote these nearest neighbors $\mathbf{x}'_1, \dots, \mathbf{x}'_5$. Next, we predict the future of $\mathbf{x}$ by computing the d-barycenter of $(\mathbf{x}'_k)_{k=1..5}$ while keeping the first t_0 observations of $\mathbf{x}$ fixed. The full pipeline is illustrated in Figure 5. The predictions obtained for the example shown in Figure 5 are illustrated in Figure 6. While ℓ_2 (Euclidean mean, DBA) based method clearly fail to identify neighbors in the same class ("a"), OT based methods do not. Moreover, thanks to temporal variability, STA provides a more accurate prediction of the remaining time points than OT alone.

Figure 7 shows a more quantitative comparison where we evaluate the accuracy of the predictions for all samples in the dataset with the ℓ_2 and the EMD (Earth mover distance) metrics averaged across the 8 predicted time points. To compute the EMD scores, we normalize all images so that their values add up to 1 and define EMD with the Euclidean quadratic cost between pixel coordinates. On both metrics and for all letters, STA outperforms the other loss functions. The prediction error of DBA in this experiment was so high that

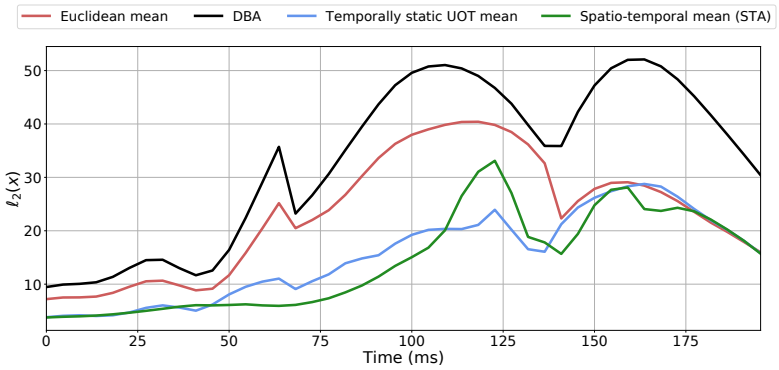

Fig. 4. ℓ_2 norm (across space) of the barycenters for the 4 methods shown in Figure 3. Euclidian average and DBA output temporally smoothed signals. On the contrary STA returns two clear peaks in the signal at 125 ms and around 160 ms, suggesting that it has correctly identified the two transient evoked responses elicited by the visual presentation of faces to the different subjects.

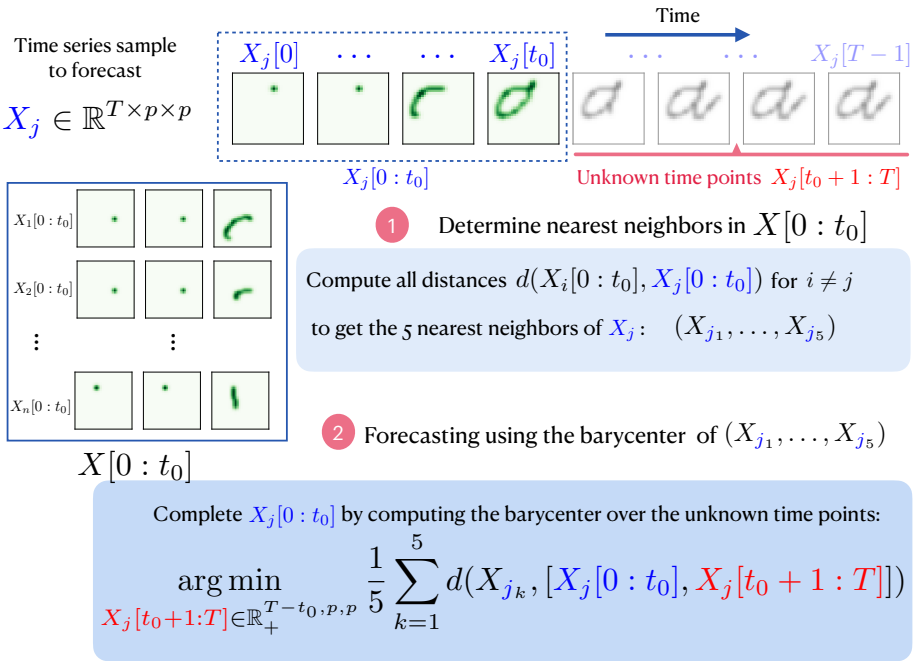

Fig. 5. Sketch explaining the forecasting pipeline used with the handwritten letters experiment.

the nuance between the other methods was no longer visible. For this reason, we defer the performance of DBA to the appendix.

Fig. 6. Forecasts of a handwritten letter time series. The green time points are fixed and considered known for all models. Blue observations are predicted. As expected ℓ_2 based methods fail to identify neighbors in the same class.

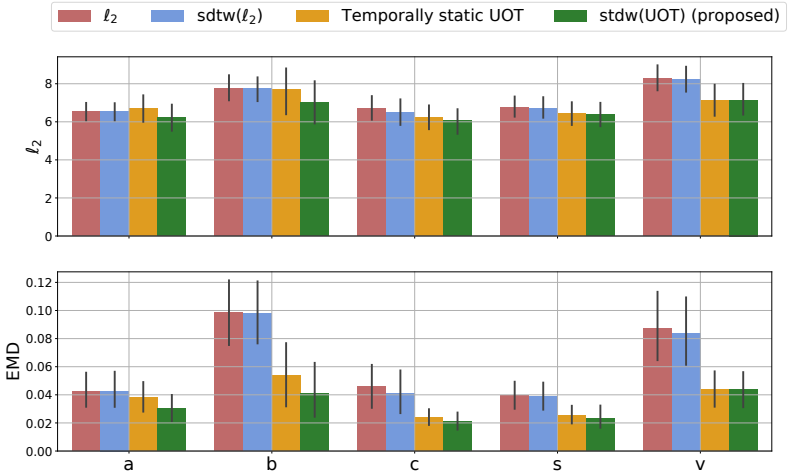

Fig. 7. Mean prediction scores computed on the unknown halves of the time series for each letter. The EMD score is computed between each true image and its prediction after normalization to the probability simplex. The ground metric is the Euclidean distance between pixel locations, normalized so that EMD is within [0-1].

6. Conclusion

Spatio-temporal data can differ in amplitude and in spatio-temporal structure. To average such data we proposed a combination of Soft-DTW and optimal transport. Our main contributions are twofold. First, we derive a fast optimization schedule that reduces the spatio-temporal averaging to a sequence of temporally weighted spatial averages. Second, combined with entropic unbalanced OT as a spatial mean, we show that the debiased version can also be used to compute barycenters using a generalized version of Sinkhorn's algorithm. The performance of our experiments on simulations and real data confirm our findings and show that our method can identify meaningful spatio-temporal barycenters.

Acknowledgements

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

Supplementart materials

Proof of Proposition 4.

The central (or diagonal) Delannoy numbers D_m verify the 2-stages recursion equation for any $m \geq 2$ (Stanley, 2011):

$$mD_{m+1} = (6m-3)D_m - (m-1)D_{m-1} \tag{44}$$

We are going to prove both inequalities by induction.

Inequality (8) For $m = 1$, we have $D_2 = 3 \leq 2 + \frac{4}{3}\sqrt{2} = \frac{2}{3}c^2 = \frac{2}{3}c^2 D_1$. Assume that (9) holds for some $m \geq 2$. From (44) and the induction assumption:

$$\begin{aligned}
(m+1)D_{m+2} &= (6m+3)D_{m+1} - mD_m && (45)\\
&\leq (6m+3)D_{m+1} - m\frac{m+\frac{1}{2}}{c^2 m}D_{m+1} && (46)\\
&\leq (6m+3-\frac{m+\frac{1}{2}}{c^2})D_{m+1} && (47)\\
&= \frac{(6c^2-1)m + \frac{6c^2-1}{2}}{c^2}D_{m+1} && (48)\\
&= c^2(m+\frac{1}{2})D_{m+1} && (49)
\end{aligned}$$

where we used the fact that $1/c^2 = \frac{1}{3+2\sqrt{2}} = 3 - 2\sqrt{2} = 6 - c^2$, hence $6c^2 - 1 = c^4$. Therefore:

$$\frac{D_{m+2}}{D_{m+1}} \leq c^2 \frac{m+\frac{1}{2}}{m+1} \quad .$$

To conclude, it suffices to show that for all $m \geq 5$:

$$\frac{m+\frac{1}{2}}{m+1} \leq \frac{m+1}{m+\frac{3}{2}} \tag{50}$$

which is equivalent to:

$$\left(m+\frac{1}{2}\right)\left(m+\frac{3}{2}\right) \leq (m+1)^2 \tag{51}$$

$$\Leftrightarrow m^2+2m+\frac{3}{4} \leq m^2+2m+1 \tag{52}$$

$$\Leftrightarrow \frac{3}{4} \leq 1 \tag{53}$$

Inequality (9) For $m=1$, we have with numerical evaluation $\frac{D_2}{D_1} - c^2\frac{2}{5} > 0$. Assume that (9) holds for some $m \geq 1$. From (44) and the induction assumption:

$$(m+1)D_{m+2} = (6m+3)D_{m+1} - mD_m \tag{54}$$

$$\geq (6m+3)D_{m+1} - m\frac{m+\frac{1}{2}+\frac{1}{2m}}{c^2 m}D_{m+1} \tag{55}$$

$$= (6m+3-\frac{m+\frac{1}{2}+\frac{1}{2m}}{c^2})D_{m+1} \tag{56}$$

$$= \frac{(6c^2-1)m+3c^2-\frac{1}{2}-\frac{1}{2m}}{c^2}D_{m+1} \tag{57}$$

$$= c^2(m+\frac{\frac{c^4}{2}-\frac{1}{2m}}{c^4})D_{m+1} \tag{58}$$

$$\tag{59}$$

where we used the fact that $1/c^2 = \frac{1}{3+2\sqrt{2}} = 3-2\sqrt{2} = 6-c^2$, hence $6c^2-1=c^4$. Therefore:

$$\frac{D_{m+2}}{D_{m+1}} \geq c^2(m+\frac{1}{2}-\frac{1}{2c^4 m}) \ .$$

To conclude, it suffices to show that for all $m \geq 2$:

$$\frac{m+\frac{1}{2}-\frac{1}{2c^4 m}}{m+1} \geq \frac{m+1}{m+\frac{3}{2}+\frac{1}{2m+2}} \tag{60}$$

which is equivalent to:

$$\left(m+\frac{1}{2}-\frac{1}{2c^4 m}\right)\left(m+\frac{3}{2}+\frac{1}{2m+2}\right) \geq (m+1)^2$$

$$\Leftrightarrow -\frac{1}{2c^4 m}\left(m+\frac{3}{2}\right)+\frac{1}{2m+2}\left(m+\frac{1}{2}\right)-\frac{1}{4c^4 m(m+1)}+\frac{3}{4} \geq 1 \tag{61}$$

$$\Leftrightarrow -(m+1)\left(m+\frac{3}{2}\right)+c^4 m\left(m+\frac{1}{2}\right)-\frac{1}{2}-\frac{c^4 m(m+1)}{2} \geq 0 \tag{62}$$

$$\Leftrightarrow (\frac{1}{2}c^4-1)m^2-2m-2 \geq 0 \ . \tag{63}$$

Which holds for $m \geq 1$. ■

Proof of Corollary 5.

Combining both inequalities of proposition 4 leads to:

$$\begin{aligned}
&\prod_{k=m}^{T-1} c^2 \frac{k}{k+\frac{1}{2}+\frac{1}{2k}} \leq \frac{D_T}{D_m} \leq \prod_{k=m}^{T-1} c^2 \frac{k}{k+\frac{1}{2}} \\
\Leftrightarrow &\prod_{k=m}^{T-1} \frac{k+\frac{1}{2}}{k} \leq c^{2(T-m)} \frac{D_m}{D_T} \leq \prod_{k=m}^{T-1} \frac{k+\frac{1}{2}+\frac{1}{2k}}{k} \\
\Leftrightarrow &\prod_{k=m}^{T-1} \frac{k+\frac{1}{2}}{k} \leq c^{2(T-m)} \frac{D_m}{D_T} \leq \prod_{k=m}^{T-1} \frac{k+\frac{1}{2}+\frac{1}{2k}}{k} \\
\Leftrightarrow &\exp\left[\sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}\right)\right] \leq c^{2(T-m)} \frac{D_m}{D_T} \leq \exp\left[\sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}+\frac{1}{2k^2}\right)\right]
\end{aligned}$$

Using the inequalities $\frac{x}{1+x} \leq \log(1+x) \leq x$ which holds for $x \geq -1$, on one hand:

$$\begin{aligned}
&\sum_{k=m}^{T-1} \frac{1}{1+2k} \leq \sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}\right) \\
\Rightarrow &\sum_{k=2m+1}^{2T} \frac{1}{k} - \sum_{k=m+1}^{T} \frac{1}{2k} \leq \sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}\right) \\
\Rightarrow &\sum_{k=1}^{2T} \frac{1}{k} - \sum_{k=1}^{2m} \frac{1}{k} - \sum_{k=1}^{T} \frac{1}{2k} + \sum_{k=1}^{m} \frac{1}{2k} \leq \sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}\right)
\end{aligned}$$

and on the other hand, for $m \geq 3$:

$$2\sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}+\frac{1}{2k^2}\right) \leq \sum_{k=1}^{T-1} \frac{1}{k} - \sum_{k=1}^{m-1} \frac{1}{k} + \frac{\pi^2}{6} - \frac{5}{4} \ .$$

Finally, using –larger versions for the sake of simplicity of – the classical bounds of the Harmonic series (Chen and Qi, 2003):

$$\log(n) + \gamma + \frac{1}{2n+1} \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log(n) + \gamma + \frac{1}{2n-1} \ , \tag{64}$$

which leads to:

$$\log(\frac{T-1}{m-1}) + \frac{1}{2T-1} - \frac{1}{2m-3} \leq \sum_{k=m}^{T-1} \frac{1}{k} \leq \log(\frac{T-1}{m-1}) + \frac{1}{2T-3} - \frac{1}{2m-1} \ , \tag{65}$$

it holds:

$$\begin{aligned}
2\sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}+\frac{1}{2k^2}\right) &\leq \log\left(\frac{T-1}{m-1}\right) + \frac{1}{2T-3} - \frac{1}{2m-1} + \frac{\pi^2}{6} - \frac{5}{4} \\
&\leq \log\left(\frac{T-1}{m-1}\right) + \frac{\pi^2}{6} - \frac{5}{4}
\end{aligned}$$

and similarly:

$$\sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}\right) \geq \log\left(\frac{2T}{2m}\right) + \frac{1}{2}\log\left(\frac{m}{T}\right) + \frac{1}{4T+1} - \frac{1}{4m-1} + \frac{1}{4m+2} - \frac{1}{4T-2}$$
$$\geq \frac{1}{2}\log\left(\frac{T}{m}\right) - \frac{1}{2}$$

Taking the exponential ends the proof ■.

6.1 Proof of proposition 8

For empty time series, we adopt the convention $D_{0,i} = D_{i,0} = 1$ for any integer i. First let's upper bound $\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y})$. Notice that since $\mathbf{x},\mathbf{y}$ are Dirac time series, the elements of the distance matrix $\boldsymbol{\Delta}(\mathbf{x},\mathbf{y})$ are either equal to 0 or r. Therefore, the cost of any path A given by $\langle A, \boldsymbol{\Delta}(\mathbf{x},\mathbf{y})\rangle$ can be written as qr for some $q \in \mathbb{N}$. More specifically, q corresponds to the number of times the path A meets the non-zero elements of $\boldsymbol{\Delta}(\mathbf{x},\mathbf{y})$. Therefore, denoting the number of feasible alignments corresponding to each q by $M_q(\mathbf{x},\mathbf{y})$ it holds:

$$\sum_{A\in\mathcal{A}_{T,T}} e^{-\frac{\langle A,\boldsymbol{\Delta}(\mathbf{x},\mathbf{y})\rangle}{\beta}} = \sum_{q=0}^{T+k} M_q(\mathbf{x},\mathbf{y}) e^{-\frac{qr}{\beta}} = M_0(\mathbf{x},\mathbf{y}) + e^{-\frac{r}{\beta}} \sum_{q=1}^{T+k} M_k(\mathbf{x},\mathbf{y}) e^{-\frac{r(q-1)}{\beta}} \tag{66}$$

Therefore:

$$\begin{aligned}\sum_{A\in\mathcal{A}_{T,T}} e^{-\frac{\langle A,\boldsymbol{\Delta}(\mathbf{x},\mathbf{y})\rangle}{\beta}} &\leq M_0(\mathbf{x},\mathbf{y}) + e^{-\frac{r}{\beta}} \sum_{q=1}^{T+k} M_k(\mathbf{x},\mathbf{y}) \\ &= M_0(\mathbf{x},\mathbf{y}) + e^{-\frac{r}{\beta}}(D_T - M_0(\mathbf{x},\mathbf{y})) \\ &= (1-\lambda_\beta) M_0(\mathbf{x},\mathbf{y}) + \lambda_\beta D_T\end{aligned}$$

and similarly:

$$\sum_{A\in\mathcal{A}_{T,T}} e^{-\frac{\langle A,\boldsymbol{\Delta}(\mathbf{x},\mathbf{x})\rangle}{\beta}} \geq M_0(\mathbf{x},\mathbf{x})$$

where $M_0(\mathbf{x},\mathbf{y}) = D_{t^\star-1,t^\star-1+k} D_{T-t^\star,T-t^\star-k}$ and $M_0(\mathbf{x},\mathbf{x}) = D_{t^\star-1} D_{T-t^\star}$. Therefore, combining both inequalities after introducing $-\log$ leads to:

$$\begin{aligned}\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y}) - \mathbf{dtw}_\beta(\mathbf{x},\mathbf{x}) &\geq -\beta \log\left(\frac{M_0(\mathbf{x},\mathbf{y})(1-\lambda_\beta) + \lambda_\beta D_T}{M_0(\mathbf{x},\mathbf{x})}\right) \\ &= -\beta \log\left(\frac{D_{t^\star-1,t^\star-1+k} D_{T-t^\star,T-t^\star-k}}{D_{t^\star-1} D_{T-t^\star}}(1-\lambda_\beta) + \lambda_\beta \frac{D_T}{D_{t^\star-1} D_{T-t^\star}}\right)\end{aligned}$$

On one hand applying Proposition 6 with $m = t^\star - 1$ and $m' = T - m - 1$ provides:

$$\frac{D_{t^\star-1,t^\star-1+k} D_{T-t^\star,T-t^\star-k}}{D_{t^\star-1} D_{T-t^\star}} \leq e^{-P(k)}$$

and on the other hand, using Corollary 5 we can get the upper bound constant H. For $t^\star \geq 1$:

$$\frac{D_T}{D_{T-t^\star}} \leq c^{2t^\star} \sqrt{\frac{(T-t^\star)e}{T}}$$

and if $t^\star - 1 \geq 2$:

$$\frac{D_2}{D_{t^\star-1}} \leq \frac{e^{\frac{\pi^2}{6}}}{c^{2(t^\star-3)}} \sqrt{t^\star - 2}$$

Combining the two leads to:

$$\frac{D_T}{D_{t^\star-1} D_{T-t^\star}} \leq \frac{c^6 e^{\frac{\pi^2}{12} - \frac{5}{8}}}{3} \sqrt{\frac{e(T-t^\star)(t^\star-2)}{T}}$$

Maximizing the upper bound with respect to $t^\star$ leads to the maximizer $t^\star = \frac{T}{2}+1$. Substituting shows that: Finally:

$$\frac{D_T}{D_{t^\star-1} D_{T-t^\star}} \leq \frac{c^6 e^{\frac{\pi^2}{6} - \frac{5}{8}}}{6} \sqrt{eT}$$

Multiplying by $1 - \lambda_\beta \geq 0$, adding $\lambda_\beta H$ and applying $-\log$ ends the proof ■.

Proof of proposition 9

It is straightforward to see that $\lim_{k \to +\infty} \mathrm{LB}_\beta(k) = -\beta \log(\lambda_\beta H)$. Therefore on one hand:

$$\begin{aligned}
&\frac{lim}{k \to +\infty} \frac{\mathrm{LB}_\beta(k) - \mathrm{LB}_\beta(k_{\max})}{\beta} \leq \eta \\
\Leftrightarrow &\log\left(e^{-P(k_{\max})}(1-\lambda_\beta) + \lambda_\beta H\right) - \log(\lambda_\beta H) \leq \eta \\
\Leftrightarrow &e^{-P(k_{\max})}(1-\lambda_\beta) + \lambda_\beta H \leq e^\eta \lambda_\beta H \\
\Leftrightarrow &\lambda_\beta \geq \frac{e^{-P(k_{\max})}}{(e^\eta - 1)H + e^{-P(k_{\max})}}
\end{aligned}$$

and on the other hand:

$$\begin{aligned}
&\beta \geq \frac{r}{P(k_{\max}) + \log\left((e^\eta - 1)H\right)} \\
\Rightarrow &-\frac{r}{\beta} \geq -P(k_{\max}) - \log\left((e^\eta - 1)H\right) \\
\Rightarrow &\lambda_\beta \geq \frac{e^{-P(k_{\max})}}{(e^\eta - 1)H} \geq \frac{e^{-P(k_{\max})}}{(e^\eta - 1)H + e^{-P(k_{\max})}}
\end{aligned}$$

Thus the upper bound in (18) holds. The lower bound follows from the positivity of $e^{-P(k)}(1-\lambda_\beta)$ ■.

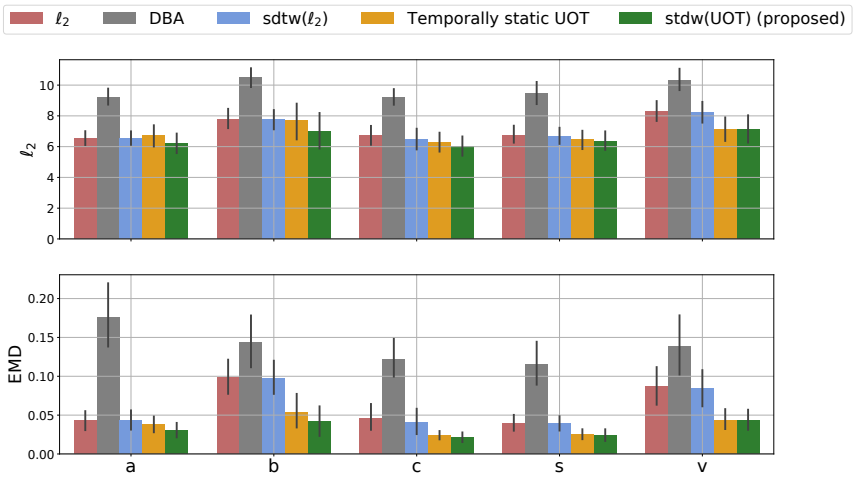

Fig. 8. Mean prediction scores computed on the unknown halves of the time series for each letter. The EMD score is computed between each true image and its prediction after normalization to the probability simplex. The ground metric is the Euclidean distance between pixel locations, normalized so that EMD is within [0-1]. The poor performance of DBA is most likely due to its (1) high sensitivity with respect to initialization and (2) its blindness to temporal shifts.

Additional experimental results

References

Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 8(6): 551–565, 1996.

Chao-Ping Chen and Feng Qi. The best bounds of harmonic sequence. *arXiv preprint math/0306233*, 2003.

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

Samuel Cohen, Giulia Luise, Alexander Terenin, Brandon Amos, and Marc Deisenroth. Aligning time series on incomparable spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 1036–1044. PMLR, 2021.

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

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

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

Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60 (4):941–965, 2018.

Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–413. IEEE, 2007.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. *ArXiv*, abs/1803.10081, 2018.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. pages 291–299, 2017.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 10 2018.

Alexandre Gramfort, Theodore Papadopoulo, Sylvain Baillet, and Maureen Clerc. Tracking cortical activity from M/EEG using graph cuts with spatiotemporal constraints. *NeuroImage*, 54(3):1930 – 1941, 2011. ISSN 1053-8119.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013. ISSN 1662-453X. doi: 10.3389/fnins.2013.00267. URL https://www.frontiersin.org/article/10.3389/fnins.2013.00267.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. 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, 16–19 Apr 2019.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In *Proceedings of the Twenty-third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 03–05 Jun 2020a.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4692–4701. PMLR, 13–18 Jul 2020b. URL http://proceedings.mlr.press/v119/janati20a.html.

Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17 (11):4302–4311, 1997. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.17-11-04302.1997. URL https://www.jneurosci.org/content/17/11/4302.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

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

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

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, February 1978. ISSN 0096-3518.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Scott D. Slotnick, Stanley A. Klein, Thom Carney, Erich Sutter, and Shahram Dastmalchi. Using multi-stimulus vep source localization to obtain a retinotopic map of human primary visual cortex. *Clinical Neurophysiology*, 110(10):1793 – 1800, 1999. ISSN 1388-2457. doi: https://doi.org/10.1016/S1388-2457(99)00135-2. URL http://www.sciencedirect.com/science/article/pii/S1388245799001352.

J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, July 2015. ISSN 0730-0301.

Richard P. Stanley. *Enumerative Combinatorics: Volume 1*. Cambridge University Press, New York, NY, USA, 2nd edition, 2011. ISBN 1107602629, 9781107602625.

Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation l(p) distance for signal analysis. *Journal of mathematical imaging and vision*, 59(2):187–210, 10 2017.

Titouan Vayer, Laetitia Chapel, Nicolas Courty, Rémi Flamary, Yann Soullard, and Romain Tavenard. Time series alignment with global invariances. *arXiv preprint arXiv:2002.03848*, 2020.

D.G. Wakeman and R.N.A. Henson. A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2(150001), 2015.

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

Tal Yarkoni. Neurosynth core tools v0.3.1, May 2014. URL https://doi.org/10.5281/zenodo.9925.