PT$\mathbb{L}^p$: Partial Transport $\mathbb{L}^p$ Distances

Xinran Liu[*]
Department of Computer Science
Vanderbilt University
Nashville, TN 37235
xinran.liu@vanderbilt.edu

Yikun Bai[*]
Department of Computer Science
Vanderbilt University
Nashville, TN 37235
yikun.bai@vanderbilt.edu

Huy Tran
Department of Computer Science
Vanderbilt University
Nashville, TN 37235
huy.tran@vanderbilt.edu

Zhanqi Zhu
Department of Computer Science
Vanderbilt University
Nashville, TN 37235
zhanqi.zhu@vanderbilt.edu

Matthew Thorpe
Department of Mathematics
The University of Manchester
Manchester, UK
matthew.thorpe-2@manchester.ac.uk

Soheil Kolouri
Department of Computer Science
Vanderbilt University
Nashville, TN 37235
soheil.kolouri@vanderbilt.edu

Abstract

Optimal transport and its related problems, including optimal partial transport, have proven to be valuable tools in machine learning for computing meaningful distances between probability or positive measures. This success has led to a growing interest in defining transport-based distances that allow for comparing signed measures and, more generally, multi-channeled signals. Transport L^p distances are notable extensions of the optimal transport framework to signed and possibly multi-channeled signals. In this paper, we introduce partial transport L^p distances as a new family of metrics for comparing generic signals, benefiting from the robustness of partial transport distances. We provide theoretical background such as the existence of optimal plans and the behavior of the distance in various limits. Furthermore, we introduce the sliced variation of these distances, which allows for faster comparison of generic signals. Finally, we demonstrate the application of the proposed distances in signal class separability and nearest neighbor classification.

1 Introduction

At the heart of Machine Learning (ML) lies the ability to measure similarities or differences between signals existing in different domains, such as temporal, spatial, spatiotemporal grids, or even graphs in a broader sense. The effectiveness of any ML model depends significantly on the discriminatory power of the metrics it employs. Several criteria are desired when quantifying dissimilarities among diverse signals, including: 1) the ability to compare signals with varying lengths, 2) adherence to the inherent structure and geometry of the signals' domain, 3) being invariant to local deformation and symmetries, 4) computational efficiency, and 5) differentiability. In recent literature, significant efforts have been dedicated to addressing these challenges. Prominent examples include the Dynamic

[*]Equal contribution

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

Time Warping (DTW) [1] technique and its numerous extensions [2, 3, 4, 5, 6], as well as more recent methods based on optimal transport principles [7, 8, 9, 10].

Optimal Transport. Optimal transport (OT) has gained recognition as a powerful tool for quantifying dissimilarities between probability measures, finding broad applications in data science, statistics, machine learning, signal processing, and computer vision [11, 12]. The dissimilarity metrics derived from OT theory define a robust geometric framework for comparing probability measures, exhibiting desirable properties such as a weak Riemannian structure [13], the concept of barycenters [14], and parameterized geodesics [15]. However, it is important to note that OT has limitations when it comes to comparing general multi-channel signals. OT is specifically applicable to non-negative measures with equal total mass, restricting its use to signals that meet specific criteria: 1) single-channel representation, 2) non-negativity, and 3) integration to a common constant, such as unity for probability measures. In cases where signals do not fulfill these criteria, normalization or alternative methods are required for meaningful comparison using OT.

Unbalanced and Optimal Partial Transport. Comparing non-negative measures with varying total amounts of mass is a common requirement in physical-world applications. In such scenarios, it is necessary to find partial correspondences or overlaps between two non-negative measures and compare them based on their respective corresponding and non-corresponding parts. Recent research has thus focused on extensions of the OT problem that enable the comparison of non-negative measures with unequal mass. The Hellinger-Kantorovich distance [16, 17], optimal partial transport (OPT) problem [18, 19, 20, 21], Kantorovich-Rubinstein norm [22, 23] and unnormalized optimal transport [24, 25] are some of the variants that fall under the category of "unbalanced optimal transport" [16, 17]. These methods provide effective solutions for comparing non-negative measures in scenarios where the total amount of mass varies. It is important to note that although the unbalanced optimal transport methods have advanced the capabilities of comparing non-negative measures with unequal mass, they still cannot be used to compare multi-channel signals or signed signals.

Transport-Based Comparison of Generic Signals. Recent studies have proposed extensions of the Optimal Transport (OT) framework to compare multi-channel signals that may include negative values, while still harnessing the benefits of OT. For example, Su and Hua [8] introduced the Order-preserving Wasserstein distance, which computes the OT problem between elements of sequences while ensuring temporal consistency through regularization of the transportation plan. A more rigorous treatment of the problem was proposed in [7] that led to the so-called Transportation L^p (TL^p) distances. In short, to compare two signals f and g, TL^p uses the OT distance between their corresponding measures, e.g., the Lebesgue measure, raised onto the graphs of the signals (See Section 3). It is worth noting that within the TL^p framework, a component of the ground cost that's defined over the signal domain, effectively addresses temporal inconsistency. This aligns with the objective set forth in [8]. Later, Zhang et al. [10] utilized a similar approach to TL^p while adding entropy regularization [26] and introduced Time Adaptive OT (TAOT). Lastly, in Spatio-Temporal Alignments, Janati et al. [9] combine OT with softDTW. They utilized regularized OT to capture spatial differences between time samples and employed softDTW for temporal alignment costs.

Contributions. In this paper, we tackle the problem of comparing multi-channel signals using transport-based methods and present a new family of metrics, denoted as PTL^p, based on the optimal partial transport framework. Our approach is motivated by the realization that while TL^p distances allow for the comparison of general signals, they require complete correspondences between input signals, which limits their applicability to real-world signals that often exhibit partial correspondences. Our specific contributions are: 1) introducing a new family of metrics based on optimal partial transport for comparing multi-channel signals, 2) providing theoretical results on existence of the partial transport plan in the proposed metric, as well as the behavior of the distance in various limits, 3) providing the sliced variation of the proposed metric with significant computational benefits, and 4) demonstrating the robust performance of the proposed metric on nearest neighbor classification in comparison with various recent baselines.

2 Background - Optimal (Partial) Transport and Their Sliced Variations

Optimal Transport. The OT problem in the Kantorovich formulation [27] is defined for two probability measures μ and ν in $\mathcal{P}(\Omega)$, and a lower semi-continuous cost function $c : \Omega^2 \to \mathbb{R}+$ by:

$$\mathrm{OT}_c(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega^2} c(x, y)\, \mathrm{d}\gamma(x, y), \tag{1}$$

Here, $\Pi(\mu, \nu)$ is the set of all joint probability measures whose marginals are μ and ν. We represent this by $\pi_{1\#}\gamma = \mu$ and $\pi_{2\#}\gamma = \nu$, where π_1 and π_2 are the canonical projection maps. If $c(x, y)$ is a p-th power of a metric, then the p-th root of the resulting optimal value is known as the p-Wasserstein distance. This distance is a metric in $\mathcal{P}_p(\Omega)$. We will ignore the subscript c if it is the default cost $\| \cdot \|^p$. Please see the appendix for more details.

Optimal Partial Transport. The problem of Optimal Partial Transport (OPT) extends the concept of mass transportation to include mass destruction at the source and mass creation at the target, with corresponding penalties for such actions. More precisely, let $\mu, \nu \in \mathcal{M}_+(\Omega)$, where $\mathcal{M}_+(\Omega)$ is set of positive Radon measures defined on Ω. Let $\lambda \geq 0$ denote the penalty for mass creation or destruction. Then the OPT problem is defined as:

$$\mathrm{OPT}_{\lambda,c}(\mu, \nu) := \inf_{\gamma \in \Pi_{\leq}(\mu,\nu)} \int_{\Omega^2} c(x, y)\, \mathrm{d}\gamma(x, y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}) \tag{2}$$

where

$$\Pi_{\leq}(\mu, \nu) := \{\gamma \in \mathcal{M}_+(\Omega^2) : \pi_{1\#}\gamma \leq \mu, \pi_{2\#}\gamma \leq \nu\},$$

$\pi_{1\#}\gamma \leq \mu$ indicates that $\pi_{1\#}\gamma$ is *dominated by* μ, i.e., for any Borel set $A \subseteq \Omega$, $\pi_{1\#}\gamma(A) \leq \mu(A)$, analogously for $\pi_{2\#}\gamma \leq \nu$. The cost function $c : \Omega^2 \to \mathbb{R}$ is lower semi-continuous (generally, it is nonnegative), and $\|\mu\|_{\mathrm{TV}}$ is the total variation (and the total mass) of μ, analogously for $\|\nu\|_{\mathrm{TV}}, \|\gamma\|_{\mathrm{TV}}$. When the transportation cost $c(x, y)$ is a metric, $\mathrm{OPT}_{\lambda,c}(\cdot, \cdot)$ defines a metric on $\mathcal{M}_+(\Omega)$ (see [28, Proposition 2.10], [29, Proposition 5], [25, Section 2.1] and [30, Theorem 4]). For simplicity of notation, we drop the c in the subscript of OT and OPT.

Sliced Transport. For one-dimensional measures, i.e., when $\Omega \subseteq \mathbb{R}$, both OT and OPT problems have efficient solvers. In particular, the OT problem has a closed-form solution, and for discrete measures with M and $N \geq M$ particles, it can be solved in $\mathcal{O}(N \log(N))$. Moreover, a quadratic algorithm, $\mathcal{O}(MN)$, was recently proposed in [31] for the one-dimensional OPT problem. To extend the computational benefits of one-dimensional OT and OPT problems to d-dimensional measures, recent works utilize the idea of slicing, which is rooted in the Cramér–Wold theorem [32] and the Radon Transform from the integral geometry [33, 34]. For $\theta \in \mathbb{S}^{d-1}$, a one-dimensional slice of measure $\mu \in \mathcal{M}_+(\Omega)$ can be obtained via $\langle \theta, \cdot \rangle_{\#}\mu$ where $\langle \cdot, \cdot \rangle : \Omega^2 \to \mathbb{R}$ denotes the inner product. Then for $\mu, \nu \in \mathcal{P}_p(\Omega)$ we can define the Sliced-OT (SOT) as:

$$\mathrm{SOT}(\mu, \nu) := \int_{\mathbb{S}^{d-1}} \mathrm{OT}(\langle \theta, \cdot \rangle_{\#}\mu, \langle \theta, \cdot \rangle_{\#}\nu)\, \mathrm{d}\sigma(\theta), \tag{3}$$

where $\sigma \in \mathcal{P}(\mathbb{S}^{d-1})$ is a probability measure such that $\mathrm{supp}(\sigma) = \mathbb{S}^{d-1}$, e.g., the uniform distribution on the unit hyper-sphere. Similarly, for $\mu, \nu \in \mathcal{M}_+(\Omega)$, Sliced-OPT (SOPT) [31] can be defined as:

$$\mathrm{SOPT}_\lambda(\mu, \nu) := \int_{\mathbb{S}^{d-1}} \mathrm{OPT}_{\lambda(\theta)}(\langle \theta, \cdot \rangle_{\#}\mu, \langle \theta, \cdot \rangle_{\#}\nu)\, \mathrm{d}\sigma(\theta), \tag{4}$$

where $\lambda \in \mathrm{L}^1(\sigma; \mathbb{R}_+)$ is generally a projection dependent function. This idea has been recently extended in multiple directions [35, 36, 37, 38, 39, 40, 41, 42, 43] to define robust statistical metrics, and has found diverse applications in the ML community [44, 31, 45, 46, 47, 48, 49, 50, 51, 52].

3 Partial Transport for Multi-Channel Signals

In the previous section, we discussed the suitability of OT and OPT problems (and similarly, SOT and SOPT problems) for comparing measures μ and ν in $\mathcal{P}_p(\Omega)$ or $\mathcal{M}_+(\Omega)$, respectively. In this section, we begin by defining a transport-based distance for multi-channel signals defined on a general class of measures, following the work of Thorpe et al. [7] on Transport L^p distances. We then motivate

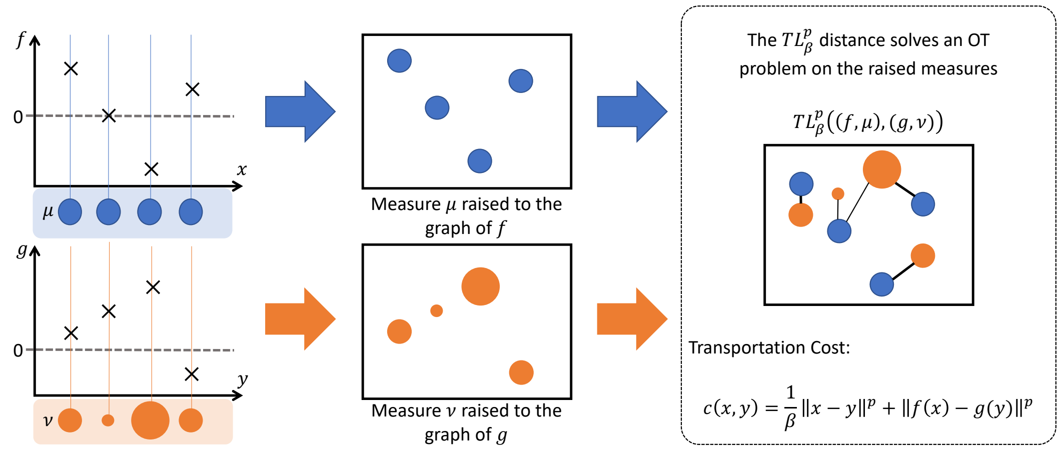

Figure 1: Illustrating the fundamental idea of TL^p distances. On the left, signals f and g are depicted along with their associated measures μ and ν. In the middle, the measures μ and ν are lifted to the graphs of f and g, respectively. On the right, the optimal transport plan is visualized, accompanied by the corresponding transportation cost.

the need for partial transportation when comparing such multi-channel signals and introduce our Partial-Transport L^p, PTL^p, distance.

Transport L^p Distances. Following [7], a multi-channel signal with k channels can be defined as the pair (f,μ) for $\mu \in \mathcal{P}_p(\Omega)$ denoting, for instance, the impulse train used for sampling the signal, and $f \in L^p(\mu;\mathbb{R}^k) := \{f : \Omega \to A \subseteq \mathbb{R}^k\}$. We denote the set of all such signals as

$$\mathcal{Q}_p(\Omega;\mathbb{R}^k) := \{(f,\mu) | \mu \in \mathcal{P}_p(\Omega), f \in \mathrm{L}^p(\mu;\mathbb{R}^k)\}.$$

We name it as the transport L^p space. The TL^p_β distance between two such k-dimensional signals (f,μ) and (g,ν) in $\mathcal{Q}_p(\Omega;\mathbb{R}^k)$ is defined as:

$$\mathrm{TL}^p_\beta((f,\mu),(g,\nu)) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\Omega^2} \Big(\frac{1}{\beta}\|x-y\|^p + \|f(x)-g(y)\|^p\Big)\,\mathrm{d}\gamma(x,y). \tag{5}$$

For any $p \in [1,\infty)$ and $\beta > 0$, the TL^p_β distance defines a proper metric on $\mathcal{Q}_p(\Omega;\mathbb{R}^k)$, and $(\mathcal{Q}_p(\Omega;\mathbb{R}^k), \mathrm{TL}^p_\beta)$ is a metric space. Intuitively, the TL^p_β measures the OT between measures μ and ν raised onto the graphs of f and g. Hence, TL^p_β solves an OT problem in the $(d+k)$-dimensional space. Figure 1 shows the core concept behind TL^p distances. Notably, the TL^p_β distance satisfies the following properties:

$$\lim_{\beta\to 0} \mathrm{TL}^p_\beta((f,\mu),(g,\nu)) = \begin{cases} \|f-g\|^p_{\mathrm{L}^p(\mu)} & \text{if } \mu=\nu \\ \infty & \text{elsewhere} \end{cases} \tag{6}$$

$$\lim_{\beta\to +\infty} \mathrm{TL}^p_\beta((f,\mu),(g,\nu)) = \mathrm{OT}(f_\#\mu, g_\#\nu) \tag{7}$$

Hence, the TL^p_β distance interpolates between the L^p distance between f, g and the p-Wasserstein distance between $f_\#\mu$ and $g_\#\nu$.

Partial Transport L^p Distances. In many real-world scenarios, it is natural for two signals to only partially match each other. Figure 2 illustrates this phenomenon. However, because TL^p distances are rooted in OT, they may sacrifice true correspondences in order to achieve a complete match between the two signals (as seen in Figure 2). To address this issue, we propose extending the definition of TL^p distances to partial transport, allowing for partial matching for signal comparison.

To do so, we first expand the definition of k-dimensional signals to be defined on positive measures rather than probability measures. Specifically, we define a signal as the pair (f,μ) where $\mu \in \mathcal{M}_+(\Omega)$

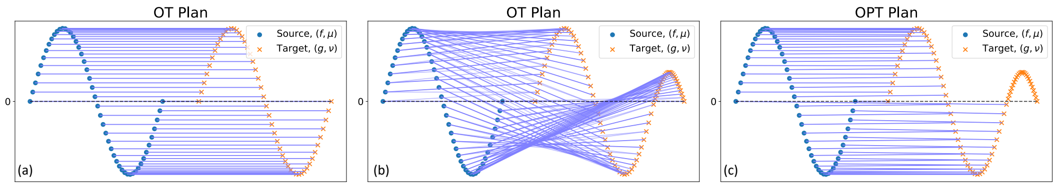

Figure 2: Highlighting the necessity of optimal partial transport. On the left, the signals exhibit complete correspondences, and OT successfully aligns the signals. In the middle, the signals display partial correspondences, and employing OT for comparison results in inaccurate correspondences. On the right, optimal partial transport (OPT) is employed, leading to the recovery of accurate partial correspondences.

and $f \in \mathrm{L}^p(f; \mathbb{R}^k)$. We denote the set of all such signals as $\mathcal{Q}_p^+(\Omega; \mathbb{R}^k)$, that is,

$$\mathcal{Q}_p^+(\Omega; \mathbb{R}^k) := \{(f, \mu) : \mu \in \mathcal{M}_+(\Omega), f \in \mathrm{L}^p(\mu; \mathbb{R}^k)\}.$$

We now propose our Partial Transport L^p (PTL^p) distance between two signals (f, μ) and (g, ν) in $\mathcal{Q}_p^+(\Omega; \mathbb{R}^k)$ as:

$$\begin{aligned}\mathrm{PTL}^p_{\beta,\lambda}((f, \mu), (g, \nu)) = \inf_{\gamma \in \Pi_{\leq}(\mu,\nu)} \int_{\Omega^2} \left(\frac{1}{\beta} \|x - y\|^p + \|f(x) - g(y)\|^p \right) \mathrm{d}\gamma(x, y) \\ + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}})\end{aligned} \tag{8}$$

Note that as opposed to the TL^p distance, the optimization in Eq. (8) is on the set of partial correspondences $\Pi_{\leq}(\mu, \nu)$. In practice, one often deals with discrete samples of k-dimensional signals. Consider f and g as two k-dimensional signals with M and N samples respectively. Let $\mu = \sum_{i=1}^M \delta_{x_i}$ and $\nu = \sum_{j=1}^N \delta_{y_j}$ denote the impulse trains used for sampling the signals (i.e., the empirical distributions), where $N, M \in \mathbb{N}$, $f_i = f(x_i)$, and $g_j = g(y_j)$. In this case, the PTL^p problem (8) can be written as:

$$\mathrm{PTL}^p_{\beta,\lambda}((f, \mu), (g, \nu)) = \sum_{\gamma \in \Pi_{\leq}(1_M, 1_N)} (\frac{1}{\beta} \|x_i - y_j\|^p + \|f_i - g_j\|^p)\gamma_{ij} + \lambda(N + M - 2|\gamma|) \tag{9}$$

where 1_M is the $M-$length vector with all 1 entries and analogously 1_N;

$$\Pi_{\leq}(1_M, 1_N) := \{\gamma \in \mathbb{R}_+^{M \times N} : \gamma 1_M \leq 1_N, \gamma^T 1_N \leq 1_M\};$$

and $|\gamma| = \sum_{ij} \gamma_{ij}$. In this case, we can further restrict the searching space of γ as the optimal γ will be induced by a 1-1 mapping.

Next, we provide some of the theoretical characteristics of PTL^p. First, the PTLP problem (8) admits a minimizer, and the optimal value defines a metric in $\mathcal{Q}_p^+(\Omega; \mathbb{R}^k)$:

Theorem 3.1. *For any $p \geq 1$, and $\lambda, \beta > 0$ there exists a minimizer for the PTL^p problem* (8)*. Furthermore, for the empirical PTL^p problem* (9)*, there exists a minimizer $\gamma \in \Pi_{\leq}(1_M, 1_N)$ that is induced by a 1-1 mapping. That is, the optimal γ satisfies $\gamma_{ij} \in \{0, 1\}$ for each i, j, and each row and column of γ contains at most one nonzero element.*

Theorem 3.2. *$(\mathcal{Q}_+(\Omega; \mathbb{R}^k), \mathrm{PTL}^p_{\beta,\lambda})$ defines a metric space.*

We refer to Section C in the appendix for the proofs of the above theorems and a detailed discussion of the PTL^p space $\mathcal{Q}_+(\Omega; \mathbb{R}^k)$.

Similar to the TL^p distance, we can also extend the definition for $\beta = 0$ and $\beta = \infty$ by the following theorem:

Theorem 3.3. *If $\lambda > 0$, we have*

$$\lim_{\beta \to 0} \mathrm{PTL}^p_{\beta,\lambda}((f, \mu), (g, \nu)) = \|f - g\|^p_{\mathrm{L}^p(\mu \wedge \nu), 2\lambda} + \lambda(\|\mu - \nu\|_{\mathrm{TV}}) \tag{10}$$

$$\lim_{\beta \to \infty} \mathrm{PTL}^p_{\beta,\lambda}((f, \mu), (g, \nu)) = \mathrm{OPT}_\lambda(f_\# \mu, g_\# \nu), \tag{11}$$

where $\mu \wedge \nu$ is the minimum of measure μ, ν,

$$\|f-g\|^p_{\mathrm{L}^p(\mu\wedge\nu),2\lambda} := \int_\Omega \|f-g\|^p \wedge 2\lambda \,\mathrm{d}(\mu\wedge\nu).$$

and $\|\mu-\nu\|_{\mathrm{TV}}$ is the total variation of the signed measure $\mu-\nu$.

See Section A in the appendix for the details of notations and Section D for the proof. Note, if we take $\lambda \to \infty$, we can recover (6), (7) by the above limits. We note that $\lambda \to 0$ is not an interesting case as it indicates zero cost for creation and destruction of mass, leading to an optimal γ of all zeros, i.e., $\mathrm{PTL}^p_{\beta,0}((\mu,f),(\nu,g)) = 0$ for all $(\mu,f),(\nu,g) \in \mathcal{Q}^p_+(\Omega;\mathbb{R}^k)$.

Sliced Extensions of TLP and PTLP. Using the connection between the TL^p distance and OT distance [7], Eq. (5) can be rewritten as

$$\mathrm{TL}^p_\beta((f,\mu),(g,\nu)) = \mathrm{OT}(\hat{\mu},\hat{\nu}) \tag{12}$$

where $\hat{\mu} = (T_{\beta,f,p})_\# \mu$ is a push-forward measure of μ by $T_{\beta,f,p}(x) = \begin{bmatrix} x\beta^{-\frac{1}{p}} \\ f(x) \end{bmatrix}$, and similarly $\hat{\nu} = (T_{\beta,g,p})_\# \nu$. Eq. (12) allows us to apply SOT method to the TL^p distance, and have the sliced-TLP distance as follows:

$$\mathrm{STL}^p_\beta((f,\mu),(g,\nu)) = \int_{\mathbb{S}^{d+k-1}} \mathrm{OT}(\theta_\#\hat{\mu},\theta_\#\hat{\nu}) d\sigma(\theta) \tag{13}$$

where $\sigma(\theta)$ is a probability measure with non-zero density on $\mathbb{S}^{d+k-1}$, for instance the uniform measure on the unit sphere. Similarly, by leveraging SOPT and the relation between PTL^p and OPT (see proposition C.3), we can define Sliced PTL^p as

$$\mathrm{SPTL}^p_{\beta,\lambda}((f,\mu),(g,\nu)) = \int_{\mathbb{S}^{d+k-1}} \mathrm{OPT}_{\lambda(\theta)}(\theta_\#\hat{\mu},\theta_\#\hat{\nu}) d\sigma(\theta) \tag{14}$$

where λ can be defined as an $L^1(\sigma,\mathbb{R}_{++})$ function of θ. Note that STL^p_β and $\mathrm{SPTL}^p_{\beta,\lambda}$ are metrics on $\mathcal{Q}(\Omega;\mathbb{R}^k)$ and $\mathcal{Q}_+(\Omega;\mathbb{R}^k)$, respectively.

Equipped with the newly proposed distances, we now demonstrate their performance in separability and nearest neighbor classification.

4 Experiments

4.1 Separability

A valid distance should be able to separate a mixture of different classes of signals. We aim to illustrate the separability of the PTL^p distance on different classes of signals in this experiment. We generate the following two classes of signals on the domain $[0,1]$: $\mathbf{S}_0$ is the class of signals with translations of one positive Gaussian bump, whereas $\mathbf{S}_1$ denotes the class of translations of signals with positive and negative Gaussian bumps. The signal examples are depicted in the left box of the top row of Figure 3. To further test the robustness, we add random blip noise, $\epsilon(t)$, to each signal in the second separability experiment. The added noise, $\epsilon(t)$, includes additive white Gaussian noise together with a randomly located blip, i.e., a sharp positive/negative transition (Figure 3, bottom left).

Figure 3 shows the 2D Multi-Dimensional Scaling (MDS) embeddings calculated from the precomputed pairwise L^p, TL^p and PTL^p distance matrices. It can be seen that PTL^p demonstrates excellent performance in effectively distinguishing between the two classes and also displays resilience against noise. However, in the presence of noise in the form of blips, the TL^p approach tends to misinterpret the noise as the predominant pattern.

4.2 Nearest Neighbor Classification

Experiment setup

To demonstrate the effectiveness of our proposed PTL^p metric and its sliced variant SPTL^p, we

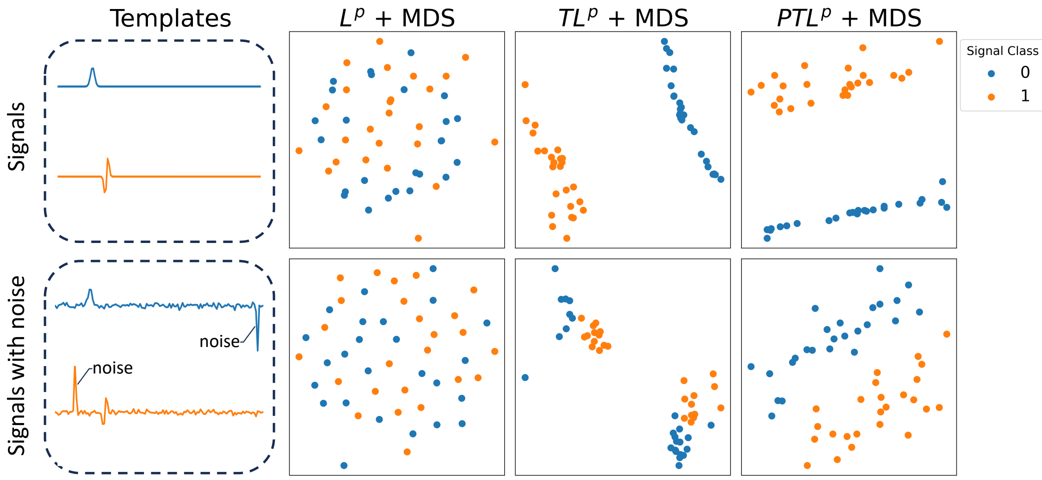

Figure 3: Visualizing manifold learning results for two classes of signals. For original signals (top row), both TL^p and PTL^p separates two classes well, but L^p fails. However, for the noisy signals (bottom row), only PTL^p shows a clear decision boundary.

evaluate the performance of these methods in the task of 1 Nearest Neighbor (1NN) classification. We compare these metrics against several other baseline approaches. The 1NN classification involves locating the closest training signal to a given test signal based on each metric or divergence, and subsequently assigning the test label according to the label of the nearest neighbor identified.

Dataset and Representation

We use three modified UCR datasets of varying lengths from [53]: Suffix, Prefix and Subsequence. The Suffix dataset is generated by simulating scenarios when sensors are activated at different times, thus may miss some observations from the start and record only suffix time series. Similarly, the Prefix dataset generator imitates the sensor behavior of stopping non-deterministically and produces only prefix time series. The Subsequence dataset contains time series that have variations on both starting and stopping time, i.e. the sensor may only capture subsequences. To represent the signals, we allocate a unit mass 1 to each discrete point of a time stamp. Consequently, differences in signal lengths will lead to variations in the total mass of the signal.

Grid search for optimal β and λ

To find the optimal β and λ for $\mathrm{PTL}^p_{\beta,\lambda}$, we perform grid search based on the 5-fold cross validation for lack of better method. We use the scikit-learn built-in GridSearchCV tools for implementation. The search range for β is set to be $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 10^3, 10^4\}$, and λ is chosen from a set of 10 evenly spaced values from 0.1 to the radius of the raised distribution on the graph of each signal.

In $\mathrm{SPTL}^p_{\beta,\lambda}$, we also need to specify the slices, i.e. θ's for 1 dimensional projections. We obtain the optimal β from $\mathrm{PTL}^p_{\beta,\lambda}$. As the amount of mass that should be transported may vary across slices, we adopt the strategy to search for the best λ for the most informative slice, and then set λ's accordingly for other slices. We set θ_0 to be the first principle component of all signals. Note that θ_0 vanishes at dimensions corresponding to $x\beta^{-\frac{1}{p}}$, but concentrates on $f(x)$ in $T_{\beta,f,p}(x) = \left[x\beta^{-\frac{1}{p}}; f(x)\right]$ (refer to Eq. (12) and Eq. (13)). Similarly, we implement grid search for best λ_{θ_0} corresponding to θ_0. Given θ_0 and λ_{θ_0}, for a specific slice θ, $\lambda_\theta = \langle\theta, \theta_0\rangle\lambda_{\theta_0}$, where $\langle\cdot,\cdot\rangle$ denotes inner product.

Results

Table 1 presents the results of nearest neighbor classification using different metrics/divergences on three subsets of the modified UCR dataset: Prefix, Subsequence, and Suffix. The table indicates that no single metric/divergence exhibits a significant advantage over others on a single dataset. However, SPTL^p achieves the best performance on two out of three datasets and performs nearly as well as the top performers on the remaining dataset, resulting in an overall win. It is worth noting that although the improvement margins are small, the computational advantage of SPTL^p and STL^p compared to other competitors (see Figure 2), make them more favorable choices in terms of efficiency.

Table 1: Nearest neighbor classification results on the modified UCR dataset [53]. For each dataset the top two performers are turned bold. The average for each subset, i.e., Prefix, Subsequence, and Suffix, as well as the total average are reported. While the overall performances are close, we note that SPTL^p and STL^p provide significantly faster solutions when accelerated by parallel computation with respect to slices.

	Method	PTL^p	SPTL^p	TL^p	STL^p	OT	SDTW $\gamma = 0.01$	SDTW $\gamma = 1$	DTW	WDTW	DDTW	L^p
Prefix Dataset	Adiac	0.23	**0.26**	0.23	0.22	0.08	0.21	0.19	0.19	0.03	**0.29**	0.25
	ArrowHead	0.49	**0.51**	0.44	0.41	0.45	0.45	0.45	0.43	0.39	**0.52**	0.25
	BeetleFly	0.65	**0.80**	0.55	0.55	0.40	0.30	0.30	0.30	0.50	**0.75**	0.55
	DSR	**0.64**	0.53	0.45	0.47	0.32	0.38	0.37	0.38	0.30	0.30	**0.60**
	DPOAG	**0.68**	**0.68**	0.65	0.64	0.59	0.65	0.63	0.65	0.42	0.42	0.55
	ECG200	**0.78**	0.74	**0.78**	0.72	0.64	0.74	0.77	0.74	0.36	0.77	0.73
	Ham	0.52	**0.57**	0.51	0.50	0.54	0.55	0.56	**0.57**	0.49	0.49	0.45
	Herring	0.63	**0.69**	0.61	0.61	0.47	**0.67**	0.66	0.66	0.59	0.59	0.56
	Lightning7	0.49	0.44	0.40	0.49	0.38	**0.55**	0.51	**0.53**	0.26	0.30	0.32
	OSULeaf	**0.45**	0.41	0.40	0.33	0.32	0.34	0.42	0.34	0.10	**0.52**	0.19
	Plane	**0.84**	**0.80**	0.69	0.68	0.48	0.68	0.72	0.67	0.14	0.14	0.50
	ShapeletSim	0.52	0.49	0.48	**0.53**	0.50	0.51	**0.57**	0.51	0.50	0.50	0.49
	SyntheticControl	0.84	0.81	0.76	0.83	0.55	**0.87**	0.84	**0.87**	0.17	0.40	0.48
	Trace	0.75	0.72	0.79	0.75	0.44	**0.81**	0.78	**0.82**	0.23	0.24	0.61
	Wine	0.50	**0.59**	0.46	0.54	**0.59**	0.44	0.46	0.46	0.50	0.56	0.50
	Average Prefix	**0.59**	**0.60**	0.55	0.55	0.45	0.54	0.55	0.54	0.33	0.45	0.47
Subsequence Dataset	Adiac	**0.17**	0.14	**0.17**	**0.17**	0.05	0.12	0.12	0.12	0.03	0.18	0.05
	ArrowHead	0.46	**0.49**	**0.51**	0.47	0.46	0.47	0.54	0.46	0.39	0.49	0.33
	BeetleFly	0.55	0.50	**0.65**	0.55	0.60	**0.65**	**0.65**	**0.65**	0.50	0.50	**0.65**
	DSR	**0.39**	**0.39**	0.34	0.34	0.34	0.43	0.38	0.35	0.30	0.31	0.29
	DPOAG	0.63	**0.65**	**0.65**	0.63	0.54	0.60	0.58	0.60	0.42	0.60	0.54
	ECG200	**0.71**	**0.69**	0.59	0.62	0.53	0.62	0.67	0.62	0.36	0.66	0.62
	Ham	0.48	**0.54**	0.49	**0.52**	0.42	0.46	0.49	0.46	0.49	0.49	0.44
	Herring	0.45	0.56	0.56	**0.58**	0.48	0.50	0.53	0.56	**0.59**	0.52	0.48
	Lightning7	0.30	0.38	0.37	**0.42**	0.29	0.37	**0.45**	0.40	0.26	0.40	0.11
	OSULeaf	**0.42**	0.38	0.32	0.31	0.24	0.34	0.41	0.34	0.10	**0.50**	0.17
	Plane	**0.57**	0.51	0.54	**0.56**	0.24	0.43	0.50	0.43	0.14	0.14	0.33
	ShapeletSim	0.49	0.46	**0.53**	**0.53**	0.51	0.53	**0.58**	0.53	0.50	0.52	0.47
	SyntheticControl	0.69	0.63	0.63	0.72	0.51	**0.76**	**0.77**	**0.76**	0.17	0.41	0.19
	Trace	0.63	0.56	0.67	0.68	0.36	**0.78**	0.64	0.74	0.23	**0.82**	0.44
	Wine	0.56	**0.67**	**0.63**	0.59	0.54	0.29	0.61	0.54	0.50	0.48	0.50
	Average Subsequence	0.50	0.50	0.51	**0.51**	0.41	0.49	**0.53**	0.50	0.33	0.47	0.37
Suffix Dataset	Adiac	0.20	0.21	**0.26**	**0.26**	0.10	0.24	0.21	0.19	0.03	**0.28**	0.06
	ArrowHead	0.45	**0.49**	0.42	0.45	0.43	0.42	0.43	0.42	0.39	**0.65**	0.29
	BeetleFly	0.50	0.60	0.55	**0.75**	0.50	0.55	**0.65**	0.55	0.50	0.50	0.55
	DSR	0.51	0.57	**0.62**	**0.62**	0.37	0.43	0.43	0.43	0.30	0.23	0.38
	DPOAG	0.59	0.61	**0.64**	0.61	0.55	**0.63**	0.59	0.62	0.42	0.42	0.43
	ECG200	0.68	**0.75**	**0.72**	0.70	0.65	0.70	0.71	0.69	0.36	0.67	0.67
	Ham	0.54	**0.57**	0.50	0.48	0.56	0.46	**0.58**	0.52	0.49	0.56	0.47
	Herring	0.56	0.56	0.52	0.52	0.47	0.56	0.50	0.48	**0.59**	**0.59**	0.58
	Lightning7	0.49	0.52	0.51	**0.58**	0.32	0.52	**0.62**	0.51	0.26	0.44	0.19
	OSULeaf	0.40	0.38	0.40	0.38	0.31	0.31	**0.41**	0.30	0.10	**0.56**	0.25
	Plane	**0.70**	**0.70**	0.64	0.63	0.48	0.67	0.68	0.67	0.14	**0.75**	0.20
	ShapeletSim	0.53	**0.56**	**0.61**	0.48	**0.56**	0.52	0.54	0.52	0.50	0.53	0.48
	SyntheticControl	0.77	0.76	0.76	0.78	0.51	**0.86**	**0.87**	**0.86**	0.17	0.44	0.10
	Trace	0.65	0.64	0.68	0.65	0.39	**0.80**	0.70	**0.76**	0.23	0.60	0.27
	Wine	0.56	**0.57**	0.56	**0.57**	0.56	**0.57**	0.54	**0.59**	0.50	0.54	**0.57**
	Average Suffix	0.54	**0.57**	0.56	**0.56**	0.45	0.55	0.56	0.54	0.33	0.52	0.37
	Average Total	**0.54**	**0.56**	0.54	0.54	0.44	0.53	0.54	0.53	0.33	0.48	0.40

4.3 Computation efficiency using Sliced PTL^p

We summarize the time complexities of all methods considered in Table 2. DTW-based methods are implemented by a dynamic programming algorithm. For DTW, soft-DTW, we use the solvers from tslearn, which are accelerated by numba. TL^p and PTL^p are solved by linear programming solvers in PythonOT, whose time complexity is cubic with respect to the length of signals in the worst case, and quadratic in practice when the measures are empirical. STL^p, SPTL^p can be accelerated by numba. For STL^p and SPTL^p, we set the number of projections to be 50. Note, the computation of STL^p and SPTL^p can be further accelerated by parallel computation with respect to slices.

Table 2: Worst case time complexities for our proposed methods and baselines. Here N denotes the length of the signals, d and k are the signal domain dimension (e.g., $d = 1$ for time series) and number of channels respectively. L is the number of slices for sliced methods. Note that DTW and its variants used in this paper share the same complexity, which is denoted by *DTW in the table.

Method	Worst-case Complexity
PTL^p	$\mathcal{O}(N^3 + N^2(d+k))$
SPTL^p	$\mathcal{O}(LN(\log(N) + (d+k)))$
TL^p	$\mathcal{O}(N^3 + N^2(d+k))$
STL^p	$\mathcal{O}(LN(\log(N) + (d+k)))$
OT	$\mathcal{O}(k(N^3 + N^2 d))$
*DTW	$\mathcal{O}(N^2 k)$
L^p	$\mathcal{O}(Nk)$

5 Conclusion

We propose partial transport L^p (PTL^p) distance as a similarity measure for generic signals. We show that PTL^p defines a metric that comes with an optimal transport plan and further characterize the behaviors of $\mathrm{PTL}^p_{\beta,\lambda}$ as β goes to various limits. We extend PTL^p to sliced partial transport L^p (SPTL^p), which is more computationally efficient. We have also demonstrated that the proposed metric is superior to other baselines in separability, and shown promising results on 1NN classification. One limitation in our methodology involves the use of a grid search for parameter selection. We leave for future studies more efficient parameter tuning of PTL^p distance in machine learning tasks.

References

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

[2] Eamonn J. Keogh and Michael J. Pazzani. *Derivative Dynamic Time Warping*, pages 1–11.

[3] Gineke A Ten Holt, Marcel JT Reinders, and Emile A Hendriks. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, volume 300, page 1, 2007.

[4] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[5] Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. Weighted dynamic time warping for time series classification. *Pattern recognition*, 44(9):2231–2240, 2011.

[6] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR, 2017.

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

[8] Bing Su and Gang Hua. Order-preserving wasserstein distance for sequence matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1057, 2017.

[9] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In *International Conference on Artificial Intelligence and Statistics*, pages 1695–1704. PMLR, 2020.

[10] Zheng Zhang, Ping Tang, and Thomas Corpetti. Time adaptive optimal transport: A framework of time series similarity measure. *IEEE Access*, 8:149764–149774, 2020.

[11] Cedric Villani. *Optimal transport: old and new*. Springer, 2009.

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

[13] Cedric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[14] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[15] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[16] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[17] Matthias Liero, Alexander Mielke, and Giuseppe Savare. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[18] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[19] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

[20] Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[21] Yikun Bai, Ivan Vladimir Medri, Rocio Diaz Martin, Rana Shahroz, and Soheil Kolouri. Linear optimal partial transport embedding. In *International Conference on Machine Learning*, pages 1492–1520. PMLR, 2023.

[22] Kevin Guittet. *Extended Kantorovich norms: a tool for optimization*. PhD thesis, INRIA, 2002.

[23] Jan Lellmann, Dirk A Lorenz, Carola Schonlieb, and Tuomo Valkonen. Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[24] Wilfrid Gangbo, Wuchen Li, Stanley Osher, and Michael Puthawala. Unnormalized optimal transport. *Journal of Computational Physics*, 399:108940, 2019.

[25] Wonjun Lee, Rongjie Lai, Wuchen Li, and Stanley Osher. Generalized unnormalized optimal transport and its fast algorithms. *Journal of Computational Physics*, 436:110041, 2021.

[26] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[27] Leonid V Kantorovich. On a problem of monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226, 1948.

[28] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[29] Benedetto Piccoli and Francesco Rossi. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[30] Yongxin Chen, Tryphon T Georgiou, Lipeng Ning, and Allen Tannenbaum. Matricial wasserstein-1 distance. *IEEE control systems letters*, 1(1):14–19, 2017.

[31] Yikun Bai, Bernard Schmitzer, Mathew Thorpe, and Soheil Kolouri. Sliced optimal partial transport. *arXiv preprint arXiv:2212.08049*, 2022.

[32] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

[33] Sigurdur Helgason et al. *Integral geometry and Radon transforms*. Springer, 2011.

[34] Lokenath Debnath and Dambaru Bhatta. *Integral transforms and their applications*. Chapman and Hall/CRC, 2016.

[35] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 32, 2019.

[36] Nadjahi et al. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

[37] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

[38] Soheil Kolouri, Kimia Nadjahi, Shahin Shahrampour, and Umut Şimşekli. Generalized sliced probability metrics. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4513–4517. IEEE, 2022.

[39] Khai Nguyen and Nhat Ho. Revisiting sliced wasserstein on images: From vectorization to convolution. *Advances in Neural Information Processing Systems*, 35:17788–17801, 2022.

[40] Khai Nguyen and Nhat Ho. Amortized projection optimization for sliced wasserstein generative models. *Advances in Neural Information Processing Systems*, 35:36985–36998, 2022.

[41] Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Minh Nguyen, and Nhat Ho. Hierarchical sliced wasserstein distance. In *The Eleventh International Conference on Learning Representations*, 2022.

[42] Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh Tan Pham. Spherical sliced-wasserstein. In *The Eleventh International Conference on Learning Representations*, 2022.

[43] Thibault Séjourné, Clément Bonet, Kilian Fatras, Kimia Nadjahi, and Nicolas Courty. Unbalanced optimal transport meets sliced-wasserstein. *arXiv preprint arXiv:2306.07176*, 2023.

[44] Soheil Kolouri, Se Rim Park, and Gustavo K Rohde. The radon cumulative distribution transform and its application to image classification. *IEEE transactions on image processing*, 25(2):920–934, 2015.

[45] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

[46] Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

[47] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

[48] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

[49] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli. Approximate bayesian computation with the sliced-wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5470–5474. IEEE, 2020.

[50] Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. *Advances in Neural Information Processing Systems*, 34:3389–3400, 2021.

[51] Yuzhe Lu, Xinran Liu, Andrea Soltoggio, and Soheil Kolouri. Slosh: Set locality sensitive hashing via sliced-wasserstein embeddings. *arXiv preprint arXiv:2112.05872*, 2021.

[52] Mohammad Shifat-E-Rabbi, Xuwang Yin, Abu Hasnat Mohammad Rubaiyat, Shiying Li, Soheil Kolouri, Akram Aldroubi, Jonathan M Nichols, and Gustavo K Rohde. Radon cumulative distribution transform subspace modeling for image classification. *Journal of Mathematical Imaging and Vision*, 63:1185–1203, 2021.

[53] Chang Wei Tan, François Petitjean, Eamonn Keogh, and Geoffrey I Webb. Time series classification for varying length series. *arXiv preprint arXiv:1910.04341*, 2019.

[54] Nicolás García Trillos and Dejan Slepčev. Continuum limit of total variation on point clouds. *Archive for rational mechanics and analysis*, 220:193–241, 2016.

[55] Florian Heinemann, Marcel Klatt, and Axel Munk. Kantorovich–rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *Applied Mathematics & Optimization*, 87(1):4, 2023.

6 Appendix

We refer to the main text for the references.

A Notation

- $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$.
- $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$.
- $\mathbb{R}^k$: the codomain of signals, where $k \geq 1$.
- Ω: unless otherwise stated is a closed subset of $\mathbb{R}^d$ where $d \geq 1$.
- $\mathcal{M}(\Omega)$, $\mathcal{M}(\Omega^2)$: the set of signed Radon measures on Ω, Ω^2 respectively.
- $\mathcal{M}_+(\Omega)$, $\mathcal{M}_+(\Omega^2)$: the set of positive Radon measures on Ω, Ω^2 respectively.
- $\mathrm{spt}(\mu)$ for $\mu \in \mathcal{M}(\Omega)$: the support of the measure μ.
- $\|\mu\|_{\mathrm{TV}}$ where $\mu \in \mathcal{M}(\Omega)$: the total variation of μ. If μ is positive, $\|\mu\|_{\mathrm{TV}} = \mu(\Omega)$.
- Weak-convergence of measures: given a sequence $\{\gamma_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(\Omega^2)$, if there exists $\gamma_0 \in \mathcal{M}(\Omega^2)$ such that

  $$\int f \,\mathrm{d}\gamma_n \to \int f \,\mathrm{d}\gamma, \forall f \in \mathrm{C}^0(\Omega^2),$$

  then we say γ_n converges weakly in measure to γ, and write

  $$\gamma_n \rightharpoonup \gamma.$$

  Note that this type of convergence is also an example of weak* convergence.
- Sequential compactness in the weak topology: given $K \subset \mathcal{M}_+(\Omega^2)$, K is said to be sequentially compact in the weak topology if for any sequence $\{\gamma_n\}_{n\in\mathbb{N}} \subset K$, there exists a further subsequence and some $\gamma \in K$ such that $\gamma_n \rightharpoonup \gamma$.
- π_1, π_2: the canonical projections on Ω^2, i.e. $\pi_1((x,y)) = x, \pi_2((x,y)) = y$.
- $f_\#\gamma$: the push-forward of the measure γ by f, i.e. $f_\#\gamma(A) = \gamma(f^{-1}(A))$ for any Borel set A.
- $\mu \wedge \nu$: minimum measure between μ and ν. Formally, for any Borel set A,

  $$\mu \wedge \nu(A) := \inf\{\mu(A_1) + \nu(A_2)\},$$

  where the infimum is taken over all possible partitions of $A = A_1 \cup A_2$. For example, if μ, ν are continuous with respect to some reference measure $\mathcal{L}$, say $\mu = f_\mu \mathcal{L}, \nu = f_\nu \mathcal{L}$, then $\mu \wedge \nu = (f_\mu \wedge f_\nu)\mathcal{L}$, where $f_\mu \wedge f_\nu$ denotes the minimum between $f_\mu(x)$ and $f_\nu(x)$ for each x.
- $\mu \leq \nu$ where $\mu, \nu \in \mathcal{M}_+(\Omega)$: if for each Borel set $A \subset \Omega$, $\mu(A) \leq \nu(A)$.
- $\Pi(\mu,\nu)$ where $\mu, \nu \in \mathcal{M}_+(\Omega)$: the set of Kantorovich plans between μ, ν, i.e. $\Pi(\mu,\nu) := \{\gamma \in \mathcal{M}_+(\Omega^2) : (\pi_1)_\#\gamma = \mu, (\pi_2)_\#\gamma = \nu\}$. Note, the set is nonempty if and only if $\mu(\Omega) = \nu(\Omega)$.
- $\Pi_\leq(\mu,\nu)$ where $\mu, \nu \in \mathcal{M}_+(\Omega)$: the set of Kantorovich plans between μ, ν for the OPT problem, i.e. $\Pi(\mu,\nu) := \{\gamma \in \mathcal{M}_+(\Omega^2) : (\pi_1)_\#\gamma \leq \mu, (\pi_2)_\#\gamma \leq \nu\}$.
- $(\mathrm{C}^0_\mathrm{b}(\Omega^2), \|\cdot\|_{\sup})$: the space of continuous and bounded functions on Ω^2. Note, the dual space of $\mathrm{C}^0_\mathrm{b}(\Omega^2)$ is the space of Radon measures $\mathcal{M}(\Omega^2)$.
- $\|\cdot\|_{\mathrm{L}^p(\mu)}, \|\cdot\|_{\mathrm{L}^p(\mu),2\lambda}$ where $\mu \in \mathcal{M}_+(\Omega)$: the L^p norm and truncated L^p norm with respect to reference measure μ. i.e.

  $$\|f\|_{\mathrm{L}^p(\mu)} = \left(\int_\Omega \|f(x)\|^p \,\mathrm{d}\mu(x)\right)^{1/p},$$

  $$\|f\|_{\mathrm{L}^p(\mu),2\lambda} = \left(\int_\Omega \|f(x)\|^p \wedge 2\lambda \,\mathrm{d}\mu(x)\right)^{1/p}$$

- $f, g : \Omega \to \mathbb{R}^k$: functions that represent the signals.
- $\beta > 0, \lambda \geq 0$: constants used in the PTL^p problem.
- $p \geq 1$: a constant which denotes the order of norm.
- $\mathcal{Q}_p^+(\Omega; \mathbb{R}^k)$: the PTL^p space, we refer to Section C for the definition.
- $D_\beta : (\Omega \times \mathbb{R}^k)^2 \to \mathbb{R}$ with

$$D_\beta^p((x, \tilde{x}), (y, \tilde{y})) = \frac{1}{\beta}\|x - y\|^p + \|\tilde{x} - \tilde{y}\|^p :$$

  the cost function in the PTL^p problem. Note, D_β defines a metric and is equivalent to ℓ^p norm on $\Omega \times \mathbb{R}^k$.
- $C(\gamma; (f, \mu), (g, \nu), \beta, \lambda) = \int_{\Omega^2} D_\beta((x, f(x)), (y, g(y))) \,\mathrm{d}\gamma(x, y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}})$: the objective function in the PTL^p problem (8).
- $C'(\gamma; (f, \mu), (g, \nu), \beta, \lambda) = \int_{\Omega^2} D_\beta((x, f(x)), (y, g(y))) \wedge 2\lambda \,\mathrm{d}\gamma(x, y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}})$: the objective function in an equivalent formulation of the PTL^p problem as defined in (18).
- $\hat{C}(\hat{\gamma}; \hat{\mu}, \hat{\nu}, D_\beta^p, \lambda) = \int_{\Omega^2} D_\beta^p((x, \tilde{x}), (y, \tilde{y}))) \,\mathrm{d}\gamma + \lambda(\|\hat{\mu}\|_{\mathrm{TV}} + \|\hat{\nu}\|_{\mathrm{TV}} - 2\|\hat{\gamma}\|_{\mathrm{TV}})$: the objective function in the OPT problem (16).
- $c_{\beta,f,g}(x, y) := \frac{1}{\beta}\|x - y\|^p + \|f(x) - g(y)\|^p$, $c_{\infty,f,g}(x, y) := \|f(x) - g(y)\|^p$: the cost function in the OPT problems (19), (24)
- $\hat{\mu} := (\mathrm{id} \times f)_\# \mu, \hat{\nu} := (\mathrm{id} \times g)_\# \nu, \hat{\gamma} := ((\mathrm{id} \times f), (\mathrm{id} \times g))_\# \gamma$, where $\mu, \nu \in \mathcal{M}_+(\Omega), \gamma \in \mathcal{M}_+(\Omega^2)$: the identification between the PTL^p space, $\mathcal{Q}_p^+(\Omega; \mathbb{R}^k)$, and a subset of $\mathcal{M}_+(\Omega \times \mathbb{R}^k)$.
- $\mathrm{OT}_c(\mu, \nu), \mathrm{OT}(\mu, \nu)$: the optimal transportation problem, where c is the cost function. The default cost is the (p-th power of the) ℓ^p norm, and we use $\mathrm{OT}(\mu, \nu)$ to denote $\mathrm{OT}_{\|\cdot\|^p}(\mu, \nu)$.
- $\mathrm{OPT}_c(\mu, \nu), \mathrm{OPT}(\mu, \nu)$: the optimal partial transportation problem, where c is the cost function. Analog to $\mathrm{OT}_c, \mathrm{OT}$.

B Preliminary Results for Radon Measures

We start by showing that weak convergence in measure preserves inequalities.

Proposition B.1. *Given a sequence $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+(\Omega)$, if $\mu_n \rightharpoonup \mu_0$ for some $\mu_0 \in \mathcal{M}(\Omega)$, then $\mu_0 \in \mathcal{M}_+(\Omega)$ and $\mu_0 \leq \mu$.*

Proof. Pick any continuous and bounded function $f \in \mathrm{C}_\mathrm{b}^0(\Omega)$, we have

$$\int f \,\mathrm{d}\mu_0 = \lim_{n \to \infty} \int_\Omega f \,\mathrm{d}\mu_n \leq \int_\Omega f \,\mathrm{d}\mu,$$

thus $\mu_0 \leq \mu$. Similarly, we have $\mu_0 \geq 0$. □

We will also make use of the closure, in the weak topology, of the space $\Pi_\leq(\mu, \nu)$, which follows similarly to the analogous result for the space $\Pi(\mu, \nu)$. A similar result has been proved when μ, ν are continuous measure, for example [18, Lemma 2.2]. For completeness we include a proof in the general case.

Proposition B.2. *The set $\Pi_\leq(\mu, \nu)$ is sequentially compact in the weak topology.*

Proof. Pick a sequence $\{\gamma_n\}_{n \in \mathbb{N}} \subset \Pi_\leq(\mu, \nu) \subset \mathcal{M}_+(\Omega^2) \subset \mathcal{M}(\Omega^2)$. We have γ_n is bounded with respect to total variation. Indeed,

$$\|\gamma_n\|_{\mathrm{TV}} = \gamma_n(\Omega^2) \leq \mu(\Omega), \quad \forall n$$

In addition, we will show γ_n is tight. Pick $\epsilon > 0$, since μ, ν are inner regular, there exists compact set $K \subset \Omega$ such that $\mu(\Omega \setminus K), \nu(\Omega \setminus K) \leq \epsilon$.

$$\gamma_n(\Omega \setminus K^2) \leq \gamma_n(\Omega \times (\Omega \setminus K) + \gamma_n((\Omega \setminus K) \times \Omega) \leq \mu(\Omega \setminus K) + \nu(\Omega \setminus K) \leq 2\epsilon.$$

Thus (γ_n) is a tight sequence.

By Prokhorov's theorem for signed measures the closure (in the weak topology) of $\Pi_{\leq}(\mu,\nu)$ is weakly sequentially compact in $\mathcal{M}(\Omega^2)$. It remains to show $\Pi_{\leq}(\mu,\nu)$ is weakly closed.

Let $\gamma_n \rightharpoonup \gamma \in \mathcal{M}(\Omega^2)$. First, we claim $\pi_{1\#}\gamma_n \rightharpoonup \pi_{1\#}\gamma, \pi_{2\#}\gamma_n \rightharpoonup \pi_{2\#}\gamma$. Indeed, pick $f \in \mathrm{C}^0_{\mathrm{b}}(\Omega)$, f can be regarded as a function in $\mathrm{C}^0_{\mathrm{b}}(\Omega^2)$ whose value is independent to the second input y. Thus, we have

$$\lim_{n\to\infty}\int_\Omega f(x)\,\mathrm{d}\pi_{1\#}\gamma_n(x) = \lim_{n\to\infty}\int_{\Omega^2} f(x)\,\mathrm{d}\gamma_n(x,y) = \int_{\Omega^2} f(x)\,\mathrm{d}\gamma(x,y) = \int_\Omega f(x)\,\mathrm{d}\pi_{1\#}\gamma(x)$$

that is $\pi_{1\#}\gamma_n \rightharpoonup \pi_{1\#}\gamma$ and similarly $\pi_{2\#}\gamma_n \rightharpoonup \pi_{2\#}\gamma$. By Lemma B.1, we have $\pi_{1\#}\gamma \leq \mu, \pi_{2\#}\gamma \leq \nu$ and $\gamma \geq 0$. □

C Relation between PTL^p and OPT

We formally introduce the PTL^p space.

Definition C.1. *Given nonempty closed $\Omega \subset \mathbb{R}^d$ the PTL^p space is defined as*

$$\mathcal{Q}_p^+(\Omega;\mathbb{R}^k) = \{(f,\mu) : \mu \in \mathcal{M}_+(\Omega), f \in \mathrm{L}^p(\mu;\mathbb{R}^k)\}.$$

The identity in $\mathcal{Q}_p^+(\Omega;\mathbb{R}^k)$ is defined as: $(f,\mu) = (g,\nu)$ if and only if $\mu = \nu$ and $f = g$, μ-a.s.

Inspired by the technique in [54], it is easy to see that the mapping $T : \mathcal{Q}_p^+(\Omega;\mathbb{R}^k) \to \mathcal{M}_+(\Omega \times \mathbb{R}^k)$ with $(f,\mu) \mapsto (\mathrm{id} \times f)_\#\mu$ defines an embedding. That is, the map T allows us to identify an element $(f,\mu) \in \mathcal{Q}_p^+(\Omega;\mathbb{R}^k)$ with a measure in the product space $\mathcal{M}_+(\Omega \times \mathbb{R}^k)$ which can be written as $(\mathrm{id} \times f)_\#\mu$.

Proposition C.2. *The mapping $T : \mathcal{Q}_p^+(\Omega;\mathbb{R}^k) \to \mathrm{Ran}(T) \subset \mathcal{M}_+(\Omega \times \mathbb{R}^k)$ is a 1-1 mapping.*

Proof. Choose distinct $(f,\mu),(g,\nu) \in \mathcal{Q}_p^+(\Omega;\mathbb{R}^k)$. Then by the identity in $\mathcal{Q}_p^+(\Omega;\mathbb{R}^k)$ as we defined above, we have one of the following: $\mu \neq \nu$ or $\mu = \nu$ and $\mu(x : f(x) \neq g(x)) > 0$.

For the first case, there exists a Borel set $A \subset \Omega$ such that $\mu(A) \neq \nu(A)$. Without loss of generality, we suppose $\mu(A) > \nu(A)$. We have

$$\begin{aligned}
T((f,\mu))(\{(x,f(x)) : x \in A\}) &= (\mathrm{id} \times f)_\#\mu(\{(x,f(x)) : x \in A\}) \\
&= \mu(A) \\
&> \nu(A) \\
&= (\mathrm{id} \times g)_\#\nu(\{(x,g(x)) : x \in A\}) \\
&\geq (\mathrm{id} \times g)_\#\nu(\{x,f(x) : x \in A, f(x) = g(x)\}) \\
&= (\mathrm{id} \times g)_\#\nu(\{(x,f(x)) : x \in A\}) \qquad (15)\\
&= T((g,\nu))(\{(x,f(x)) : x \in A\})
\end{aligned}$$

where (15) follows from the fact $(\mathrm{id} \times g)_\#\nu$ is supported on the graph of g. Thus, $T((f,\mu)) \neq T((g,\nu))$.

For the second case, there exists Borel set B, such that $f(x) \neq g(x), \forall x \in B$ and $\mu(B) > 0$. Thus, we have

$$(\mathrm{id} \times f)_\#\mu(\{(x,f(x)) : x \in B\}) = \mu(B) > 0$$

and

$$\begin{aligned}
(\mathrm{id} \times f)_\#\nu(\{(x,f(x)) : x \in B\}) &= \nu\{x : f(x) = g(x), x \in B\} \\
&= \mu\{x : f(x) = g(x), x \in B\} \\
&= 0.
\end{aligned}$$

Thus $T((f,\mu)) \neq T((g,\nu))$. Therefore T is a 1-1 mapping. □

In the space $\mathcal{M}_+(\Omega\times\mathbb{R}^k)$, we can define the following OPT problem. For $\beta\in(0,\infty)$, we define $D_\beta:(\Omega\times\mathbb{R}^k)^2\to\mathbb{R}_+$ by

$$D^p_\beta((x,\tilde{x}),(y,\tilde{y}))=\frac{1}{\beta}\|x-y\|^p+\|\tilde{x}-\tilde{y}\|^p.$$

It is straightforward to show D is metric and is equivalent to the ℓ^p metric in $\Omega\times\mathbb{R}^k$. Thus the following OPT problem defines a (p-th power of a) metric in $\mathcal{M}_+(\Omega\times\mathbb{R}^k)$ by [31, Appendix C] or [55, Theorem 2.2], where $\hat{\mu},\hat{\nu}\in\mathcal{M}_+(\Omega\times\mathbb{R}^k)$:

$$\mathrm{OPT}_{D^p_\beta,\lambda}(\hat{\mu},\hat{\nu})=\inf_{\hat{\gamma}\in\Pi_\leq(\hat{\mu},\hat{\nu})}\int_{(\Omega\times\mathbb{R}^k)^2}D^p_\beta((x,\tilde{x}),(y,\tilde{y}))\,\mathrm{d}\hat{\gamma}+\lambda(\|\hat{\mu}\|_{\mathrm{TV}}+\|\hat{\nu}\|_{\mathrm{TV}}-2\|\hat{\gamma}\|_{\mathrm{TV}}).\tag{16}$$

Similar to [54, Proposition 3.3], we will show the OPT distance (16) and the PTL^p distance (8) are equivalent.

Proposition C.3. *Choose $(f,\mu),(g,\nu)\in\mathcal{Q}^+_p(\Omega;\mathbb{R}^k)$, let $\hat{\mu}=(\mathrm{id}\times f)_\#\mu,\hat{\nu}=(\mathrm{id}\times g)_\#\nu$. Define $F:\Pi_\leq(\mu,\nu)\to\Pi_\leq(\hat{\mu},\hat{\nu})$ by*

$$\gamma\mapsto\hat{\gamma}=F(\gamma):=((\mathrm{id}\times f),(\mathrm{id}\times g))_\#\gamma.$$

Then, F is bijection. Furthermore, let $\hat{C}(\hat{\gamma};\hat{\mu},\hat{\nu},D^p_\beta,\lambda)$ and $C(\gamma;(f,\mu),(g,\nu),\beta,\lambda)$ denote the objective function in (16) *and* (8) *respectively, we have*

$$\hat{C}(\hat{\gamma};\hat{\mu},\hat{\nu},D^p_\beta,\lambda)=C(\gamma;(f,\mu),(g,\nu),\beta,\lambda).$$

Therefore,

$$\mathrm{PTL}^p_{\beta,\lambda}((f,\mu),(g,\nu))=\mathrm{OPT}_{D^p_\beta,\lambda}(\hat{\mu},\hat{\nu}),$$

Proof. First, it is straightforward that F is well defined. We start by showing that F is injective. If $\gamma_1,\gamma_2\in\Pi_\leq(\mu,\nu)$ and $\gamma_1\neq\gamma_2$ then there exists A,B such that $\gamma_1(A\times B)\neq\gamma_2(A\times B)$. Without loss of generality assume $\gamma_1(A\times B)>\gamma_2(A\times B)$. Let $\hat{\gamma}_1=F(\gamma_1)$ and $\hat{\gamma}_2=F(\gamma_2)$. Then,

$$\begin{aligned}\hat{\gamma}_1(A\times\mathbb{R}^k\times B\times\mathbb{R}^k)&=((\mathrm{id}\times f),(\mathrm{id}\times g))_\#\gamma_1(A\times\mathbb{R}^k\times B\times\mathbb{R}^k)\\&=\gamma_1\left(\{(x,y)\,:\,(x,f(x),y,g(y))\in A\times\mathbb{R}^k\times B\times\mathbb{R}^k\}\right)\\&=\gamma_1(A\times B)\\&>\gamma_2(A\times B)\\&=\gamma_2\left(\{(x,y)\,:\,(x,f(x),y,g(y))\in A\times\mathbb{R}^k\times B\times\mathbb{R}^k\}\right)\\&=((\mathrm{id}\times f),(\mathrm{id}\times g))_\#\gamma_2(A\times\mathbb{R}^k\times B\times\mathbb{R}^k)\\&=\hat{\gamma}_2(A\times\mathbb{R}^k\times B\times\mathbb{R}^k).\end{aligned}$$

So $\hat{\gamma}_1\neq\hat{\gamma}_2$, hence F is injective.

To show surjectivity, take any $\hat{\gamma}\in\hat{\Pi}_\leq(\hat{\mu},\hat{\nu})$. Take any $\hat{A},\hat{B}$ such that $0=(\hat{\mu}\times\hat{\nu})(\hat{A}\times\hat{B})=\mu(\hat{A})\nu(\hat{B})$. Then either $\hat{\mu}(\hat{A})=0$ or $\hat{\nu}(\hat{B})=0$. In the first case $\hat{\gamma}(\hat{A}\times\hat{B})\leq\hat{\gamma}(A\times(\Omega\times\mathbb{R}^k))\leq\hat{\mu}(\hat{A})=0$. Similarly, in the second case $\hat{\gamma}(\hat{A}\times\hat{B})\leq\hat{\nu}(\hat{B})=0$ and so $\mathrm{spt}(\hat{\gamma})\subseteq\mathrm{spt}(\hat{\mu}\times\hat{\nu})$. It follows in a similar way that $\mathrm{spt}(F(\gamma))\subset\mathrm{spt}(\hat{\mu}\times\hat{\nu})$ for any $\gamma\in\Pi_\leq(\mu,\nu)$.

We define $\gamma\in\mathcal{M}_+(\Omega^2)$ by $\gamma(A)=\hat{\gamma}(\{(x,f(x)),(y,g(y)):(x,y)\in A\})$. We have $\gamma\in\Pi_\leq(\mu,\nu)$. Take $\hat{A},\hat{B}\subset\Omega\times\mathbb{R}^k$ and we compare $F(\gamma)(\hat{A}\times\hat{B})$ with $\hat{\gamma}(\hat{A}\times\hat{B})$. By the previous argument we only need consider $\hat{A}\in\mathrm{spt}(\hat{\mu})$ and $\hat{B}\in\mathrm{spt}(\hat{\nu}))$. In particular, we may assume that $\hat{A}=\{(x,f(x))\,:\,x\in A\}$ and $\hat{B}=\{(y,g(y))\,:\,y\in B\}$ for some $A,B\subseteq\Omega$. Now

$$F\gamma(\hat{A}\times\hat{B})=\gamma\left(\left\{(x,y)\,:\,((x,f(x)),(y,g(y)))\in\hat{A}\times\hat{B}\right\}\right)=\gamma(A\times B)=\hat{\gamma}(\hat{A}\times\hat{B}).$$

Thus $F\gamma=\hat{\gamma}$ and so F is surjective.

For each $\gamma \in \Pi_{\leq}(\mu, \nu)$ we continue to let $\hat{\gamma} = F(\gamma)$ denote the corresponding measure in $\Pi_{\leq}(\hat{\mu}, \hat{\nu})$. We have

$$\begin{aligned}\hat{C}(\hat{\gamma}; \hat{\mu}, \hat{\nu}, D^p_\beta, \lambda) &= \int_{(\Omega\times\mathbb{R}^k)^2} D^p_\beta((x,\tilde{x}),(y,\tilde{y}))\,\mathrm{d}\hat{\gamma}(x,\tilde{x},y,\tilde{y}) + \lambda(\|\hat{\mu}\|_{\mathrm{TV}} + \|\hat{\nu}\|_{\mathrm{TV}} - 2\|\hat{\gamma}\|_{\mathrm{TV}}) \\ &= \int_{\Omega^2} \frac{1}{\beta}\|x-y\|^p + \|f(x)-g(y)\|^p\,\mathrm{d}\gamma(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}) \\ &= C(\gamma; (f,\mu),(g,\nu),\beta,\lambda) \end{aligned} \tag{17}$$

Combining with the fact that $\gamma \mapsto \hat{\gamma}$ is a bijection, we have

$$\mathrm{OPT}_\lambda(\hat{\mu}, \hat{\nu}) := \mathrm{PTL}^p_{\beta,\lambda}((f,\mu),(g,\nu))$$

which completes the proof. ☐

Remark C.4. *Proposition C.2 and C.3 imply that* $(\mathcal{Q}^+_p(\Omega), \mathrm{PTL}^p_{\lambda,\beta})$ *is a metric space when* $\beta, \lambda \in (0,\infty)$ *and therefore we can conclude Theorem 3.2.*

We now prove existence of minimizers for PTL^p problem.

Proof of Theorem 3.1. From Proposition C.3, we have that the PTL^p problem (8) admits a solution if and only if the OPT problem (16) admits a solution. By the relation between OPT and OT (see, for example, [18, section 2], or [31, Appendix B]), one can convert the OPT problem (16) to a classical OT problem defined on $(\Omega \times \mathbb{R}^k) \cup \{\hat{\infty}\}$ where $\hat{\infty}$ is an isolated point. It is lower semi-continuous and bounded from below. Thus by the classical result in optimal transport theory (e.g. [11, Theorem 4.1]), there exists an optimal transportation plan for the OPT problem (16)). Note, an equivalent way to prove the existence of a minimizer is using the direct method from the calculus of variations (compactness and lower-semi continuity implies existence of minimizers). Indeed, $\Pi_{\leq}(\mu, \nu)$ is compact in the weak topology by Proposition B.2 and $\gamma \to \hat{C}(\gamma; \hat{\mu}, \hat{\nu}, D^p_\beta, \lambda)$ is lower-semi continuous (in the sense of the weak topology). Thus the PTL^p problem (8) admits a minimizer.

For the Empirical PTL^p problem, by the relation between PTL^p and OPT as discussed in Proposition C.3 or Lemma D.1 in the next section, it suffices to show there exists an 1-1 mapping that can solve the corresponding OPT problem (see (16)). By [31, Theorem 4.1], we complete the proof. ☐

D The PTL^p Problem for Extreme β

In this section, we discuss the PTL^p problem when $\beta \to 0$ and $\beta \to \infty$. First, we prove equivalence of the PTL^p problem with a truncated version.

Lemma D.1. *There exists optimal* $\gamma \in \mathcal{M}_+(\Omega^2)$ *for the* PTL^p *problem* (8) *such that* $\gamma(S) = 0$ *where* $S := \{(x,y) : \frac{1}{\beta}\|x-y\|^p + \|f(x)-g(y)\|^p \geq 2\lambda\}$*. Therefore, the* PTL^p *problem can be defined as*

$$\begin{aligned}\mathrm{PTL}^p_{\beta,\lambda}((f,\mu),(g,\nu)) = \inf_{\gamma\in\Pi_{\leq}(\mu,\nu)} &\int_{\Omega^2} \left(\frac{1}{\beta}\|x-y\|^p + \|f(x)-g(y)\|^p\right) \wedge 2\lambda\,\mathrm{d}\gamma(x,y) \\ &+ \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}) \end{aligned} \tag{18}$$

Proof. Similar to the last section, the PTL^p problem can be written as the following OPT problem between μ, ν defined as follows:

$$\mathrm{OPT}_{c_{\beta,f,g},\lambda}(\mu,\nu) := \inf_{\gamma\in\Pi_{\leq}(\mu,\nu)} \int_{\Omega^2} c_{f,g}(x,y)\,\mathrm{d}\gamma(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}) \tag{19}$$

where the ground cost is defined as

$$c_{\beta,f,g}(x,y) := \frac{1}{\beta}\|x-y\|^p + \|f(x)-g(y)\|^p.$$

Thus, by [31, Lemma 3.2], we have that there exists an optimal γ such that $\gamma(S) = 0$ on $S = \{(x,y) : c(x,y) \geq 2\lambda\}$ and we complete the proof. ☐

Now we discuss the extreme cases $\beta = 0$ and $\beta = \infty$.

Theorem D.2. *For any positive sequence $\beta_n \to 0$, we have*

$$\lim_{n\to\infty} \mathrm{PTL}^p_{\beta,\lambda}((f,\mu),(g,\nu)) = \|f-g\|^p_{\mathrm{L}^p(\mu\wedge\nu),2\lambda} + \lambda\|\mu-\nu\|_{\mathrm{TV}} \tag{20}$$

Proof. Let $\gamma = (\mathrm{id}\times\mathrm{id})_{\#}(\mu\wedge\nu)$, then $\gamma \in \Pi_{\leq}(\mu,\nu)$ and

$$\begin{aligned}\mathrm{PTL}^p_{\beta_n,\lambda}((f,\mu),(g,\nu)) &\leq \int_{\Omega\times\Omega}\left(\frac{1}{\beta_n}\|x-y\|^p + \|f(x)-g(y)\|^p\right)\wedge 2\lambda \,\mathrm{d}\gamma(x,y) \\ &\quad + \lambda\left(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}\right) \\ &= \int_{\Omega}\|f(x)-g(x)\|^p\wedge 2\lambda\,\mathrm{d}(\mu\wedge\nu)(x) + \lambda\left(\|\mu\|_{\mathrm{TV}}+\|\nu\|_{\mathrm{TV}} - 2\|\mu\wedge\nu\|_{\mathrm{TV}}\right) \\ &= \|f-g\|^p_{\mathrm{L}^p(\mu\wedge\nu),2\lambda} + \lambda\|\mu-\nu\|_{\mathrm{TV}}.\end{aligned}$$

In the rest of the proof we will prove the coverse inequality.

Since

$$\begin{aligned}&\frac{1}{\beta_n}\left(\int_{\Omega^2}\|x-y\|^p\wedge 2\lambda\,\mathrm{d}\gamma(x,y) + \beta_n\lambda\left(\|\mu\|_{\mathrm{TV}}+\|\nu\|_{\mathrm{TV}}-2\|\gamma\|_{\mathrm{TV}}\right)\right) \\ &\qquad \leq \int_{\Omega^2}\frac{1}{\beta_n}\|x-y\|^p\wedge 2\lambda + \|f(x)-g(y)\|^p\wedge 2\lambda\,\mathrm{d}\gamma(x,y) + \lambda\left(\|\mu\|_{\mathrm{TV}}+\|\nu\|_{\mathrm{TV}}-2\|\gamma\|_{\mathrm{TV}}\right)\end{aligned}$$

then

$$\frac{1}{\beta_n}\mathrm{OPT}_{\|\cdot\|^p\wedge 2\lambda,\lambda}(\mu,\nu) \leq \mathrm{PTL}^p_{\beta_n,\lambda}((f,\mu),(g,\nu)) \leq \|f-g\|^p_{\mathrm{L}^p(\mu\wedge\nu),2\lambda} + \lambda\|\mu-\nu\|_{\mathrm{TV}}.$$

So $\mathrm{OPT}_{\|\cdot\|^p,\lambda}(\mu,\nu) = \mathrm{OPT}_{\|\cdot\|^p\wedge 2\lambda,\lambda}(\mu,\nu) \leq O(\beta_n) \to 0$.

Let $\gamma_n \in \Pi_{\leq}(\mu,\nu)$ satisfy

$$\mathrm{OPT}^p_{\|\cdot\|^p,\beta_n\lambda}(\mu,\nu) = \int_{\Omega^2}\|x-y\|^p\wedge 2\lambda\,\mathrm{d}\gamma_n(x,y) + \beta_n\lambda\left(\|\mu\|_{\mathrm{TV}}+\|\nu\|_{\mathrm{TV}}-2\|\gamma_n\|_{\mathrm{TV}}\right).$$

As $\Pi_{\leq}(\mu,\nu)$ is weakly sequentially compact then there exists a subsequence (which we relabel) and a γ_0 such that $\gamma_n \rightharpoonup \gamma_0 \in \Pi_{\leq}(\mu,\nu)$. Now, $\|\gamma_n\|_{\mathrm{TV}} = \gamma_n(\Omega^2) \to \gamma_0(\Omega^2) = \|\gamma_0\|_{\mathrm{TV}}$, and $(x,y)\mapsto \|x-y\|^p\wedge 2\lambda$ is continuous and bounded, so

$$\mathrm{OPT}_{\|\cdot\|^p\wedge 2\lambda,\beta_n\lambda}(\mu,\nu) \to \int_{\Omega^2}\|x-y\|^p\wedge 2\lambda\,\mathrm{d}\gamma_0(x,y)$$

Hence, $\int_{\Omega^2}\|x-y\|^p\,\mathrm{d}\gamma_0(x,y) = 0$ and so $x = y$ γ_0-a.e.. In particular, there exists μ_0 such that $\gamma_0 = (\mathrm{id}\times\mathrm{id})_{\#}\mu_0$ and $\mu_0 \leq \nu$, $\mu_0 \leq \nu$. We are left to show $\lim_{n\to\infty}\mathrm{PTL}^p_{\beta_n,\lambda}((f,\mu),(g,\nu)) \geq \mathrm{PTL}^p_{0,\lambda}((f,\mu),(g,\nu))$, where

$$\mathrm{PTL}^p_{0,\lambda}((f,\mu),(g,\nu)) := \inf_{\mu_0\leq\mu\wedge\nu}\int_{\Omega}\|f(x)-g(x)\|^p\wedge 2\lambda\,\mathrm{d}\mu_0(x) + \lambda(\|\mu\|_{\mathrm{TV}}+\|\nu\|_{\mathrm{TV}}-2\|\mu_0\|_{\mathrm{TV}}). \tag{21}$$

We will show that the minimizer is $\mu\wedge\nu$. Indeed, let $C'(\mu_0;(f,\mu),(g,\nu),0,\lambda)$ denote the transportation cost induced by μ_0, we have

$$\begin{aligned}&C'(\mu\wedge\nu;(f,\mu),(g,\nu),0,\lambda) - C'(\mu_0;(f,\mu),(g,\nu),0,\lambda) \\ &\quad = \int_{\Omega}\|f(x)-g(x)\|^p\wedge 2\lambda - 2\lambda\,\mathrm{d}(\mu\wedge\nu-\mu_0)(x) \\ &\quad \leq 0\end{aligned}$$

where the inequality follows since $\mu\wedge\nu-\mu_0$ is a nonnegative measure. Thus we have

$$PTL^p_{0,\lambda}((f,\mu),(g,\nu)) = \|f-g\|_{\mu\wedge\nu,2\lambda} + \lambda(\|\mu-\nu\|_{TV}). \tag{22}$$

Since $\mathrm{C}_\mathrm{b}^0(\Omega)$ is dense in $\mathrm{L}^p(\mu), \mathrm{L}^p(\nu)$, for any $\varepsilon > 0$ there exists $f_\varepsilon, g_\varepsilon \in \mathrm{C}_\mathrm{b}^0(\Omega)$ such that

$$\int_{\Omega^2} \|f(x) - f_\varepsilon(x)\|^p \,\mathrm{d}\mu(x) \leq \varepsilon, \qquad \int_{\Omega^2} \|g(y) - g_\varepsilon(y)\|^p \,\mathrm{d}\nu(y) \leq \varepsilon.$$

Now,

$$\begin{aligned}
\left(\int_{\Omega^2} \|f(x) - g(y)\|^p \,\mathrm{d}\gamma_n(x,y)\right)^{\frac{1}{p}} &\geq \left(\int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \,\mathrm{d}\gamma_n(x,y)\right)^{\frac{1}{p}} \\
&\qquad - \left(\int_{\Omega^2} \|f(x) - f_\varepsilon(x)\|^p \,\mathrm{d}\gamma_n(x,y)\right)^{\frac{1}{p}} \\
&\qquad - \left(\int_{\Omega^2} \|g_\varepsilon(y) - g(y)\|^p \,\mathrm{d}\gamma_n(x,y)\right)^{\frac{1}{p}} \\
&\geq \left(\int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \,\mathrm{d}\gamma_n(x,y)\right)^{\frac{1}{p}} - 2\varepsilon \\
&\to \left(\int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \,\mathrm{d}\gamma_0(x,y)\right)^{\frac{1}{p}} - 2\varepsilon.
\end{aligned}$$

So,

$$\begin{aligned}
\liminf_{n\to\infty} \mathrm{PTL}^p_{\beta_n,\lambda}((f,\mu),(g,\nu)) &\geq \int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \,\mathrm{d}\gamma_0(x,y) - K\varepsilon \\
&\qquad + \lambda\left(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma_0\|_{\mathrm{TV}}\right) \\
&\geq \int_{\Omega^2} \|f(x) - g(y)\|^p \,\mathrm{d}\gamma_0(x,y) - K'\varepsilon \\
&\qquad + \lambda\left(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma_0\|_{\mathrm{TV}}\right) \\
&\geq \mathrm{PTL}^p_{0,\lambda}((f,\mu),(g,\nu)) - K'\varepsilon
\end{aligned}$$

for constants K, K'. Taking $\varepsilon \to 0$ completes the proof. □

Based on the above theorem, we can extend the PTL^p distance for $\beta = 0$.

Corollary D.3. *When $\mu = \nu$, we have*

$$\mathrm{PTL}^p_{0,\lambda}((f,\mu),(g,\mu)) = \|f - g\|^p_{\mathrm{L}^p(\mu),2\lambda}.$$

The proof is straightforward.

Next, we discuss the case $\beta = \infty$.

Theorem D.4. *For any positive sequence $\beta_n \to \infty$, we have*

$$\lim_{n\to\infty} \mathrm{PTL}^p_{\beta,\lambda_n}((f,\mu),(g,\nu)) = \mathrm{OPT}_\lambda(f_\# \mu, g_\# \nu). \tag{23}$$

Proof. Without loss of generality, we can assume β_n is a monotonic increasing sequence. Note, it is straightforward to show (e.g. see Proposition C.3 by setting the $\frac{1}{\beta}$ term to be 0):

$$\begin{aligned}
\mathrm{OPT}_\lambda(f_\# \mu, g_\# \nu) &= \mathrm{OPT}_{c_{\infty,f,g},\lambda}(\mu,\nu) \\
&= \inf_{\gamma \in \Pi_{\leq}(\mu,\nu)} \int_{\Omega^2} \|f(x) - g(y)\|^p \,\mathrm{d}\gamma(x,y) + \lambda\left(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}\right)
\end{aligned} \tag{24}$$

where $c_{\infty,f,g}(x,y) := \|f(x) - g(y)\|^p$. For each n and γ, we have

$$\begin{aligned}
&\int_{\Omega^2} \frac{1}{\beta_n} \|x - y\|^p + \|f(x) - g(y)\|^p \,\mathrm{d}\gamma(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}) \\
&\geq \int_{\Omega^2} \|f(x) - g(y)\|^p \,\mathrm{d}\gamma(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma\|_{\mathrm{TV}}).
\end{aligned} \tag{25}$$

Taking the infimum on both sides and passing to the limit, we have

$$\liminf_{n\to\infty} \mathrm{PTL}^p_{\beta,\lambda_n}((f,\mu),(g,\nu)) \geq \mathrm{OPT}_{c_{f,g},\lambda}(\mu,\nu) \tag{26}$$

For the other direction, we let γ_∞ be an optimal transportation plan for $\mathrm{OPT}_{c_{f,g},\lambda}(\mu,\nu)$, then

$$\begin{aligned}
&\limsup_{n\to\infty} \mathrm{PTL}^p_{\beta,\lambda_n}((\mu,f),(\nu,g)) \\
&\leq \limsup_{n\to\infty} \int_{\Omega^2} \left(\frac{1}{\beta_n}\|x-y\|^p + \|f(x)-g(y)\|^p\right) \wedge 2\lambda \, \mathrm{d}\gamma_\infty(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma_\infty\|_{\mathrm{TV}}) \\
&= \int_{\Omega^2} \limsup_{n\to\infty} \left(\frac{1}{\beta_n}\|x-y\|^p + \|f(x)-g(y)\|^p\right) \wedge 2\lambda \, \mathrm{d}\gamma_\infty(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma_\infty\|_{\mathrm{TV}}) \\
&= \int_{\Omega^2} \|f(x)-g(y)\| \wedge 2\lambda \, \mathrm{d}\gamma_\infty(x,y) + \lambda(\|\mu\|_{\mathrm{TV}} + \|\nu\|_{\mathrm{TV}} - 2\|\gamma_\infty\|_{\mathrm{TV}})
\end{aligned}$$

where the third line follows from the Monotone convergence theorem (or Beppo Levi's lemma). Thus

$$\limsup_{n\to\infty} \mathrm{PTL}^p_{\beta,\lambda_n}((f,\mu),(g,\nu)) \leq \mathrm{OPT}_{c_{\infty,f,g},\lambda}(\mu,\nu) \tag{27}$$

Combining (26) and (27), we complete the proof. □

Remark D.5. *Combining Theorem D.2 and D.4, we prove Theorem 3.3 in the main text.*

E More 1 NN Classification Experiments

We provide additional nearest neighbor classification results in Table E.

Table 3: Additional nearest neighbor classification results on the modified UCR dataset [53]. Similarly, we highlight top two performers and provide the averages.

	Method	PTL^p	SPTL^p	TL^p	STL^p	OT	SDTW $\gamma = 0.01$	SDTW $\gamma = 1$	DTW	WDTW	DDTW	L^p
Prefix Dataset	Coffee	**0.79**	0.68	0.50	0.61	0.54	0.64	0.57	0.64	0.54	**0.71**	0.61
	ECGFiveDays	0.58	0.60	**0.63**	**0.63**	0.51	0.61	0.58	0.61	0.50	0.59	0.55
	DistalPhalanxTW	0.53	**0.54**	0.50	0.51	0.45	0.51	0.49	**0.55**	0.30	0.30	0.53
	BirdChicken	**0.75**	**0.75**	0.55	0.55	0.60	0.70	0.70	0.70	0.50	0.65	0.55
	GunPoint	0.63	**0.71**	0.65	0.66	0.67	0.65	0.69	**0.72**	0.49	0.49	0.65
	CBF	0.81	0.81	0.71	0.72	0.54	**0.86**	0.84	**0.86**	0.33	0.66	0.46
	DPOC	0.62	**0.72**	0.64	0.64	0.59	0.66	**0.67**	**0.67**	0.58	0.58	0.57
	OliveOil	**0.33**	0.30	**0.33**	**0.33**	0.30	0.27	0.23	0.23	0.17	0.17	0.13
	Symbols	**0.64**	0.61	0.42	0.56	0.32	**0.64**	**0.64**	0.62	0.16	0.56	0.47
	FaceFour	**0.66**	0.53	0.56	**0.59**	0.25	0.48	0.59	0.49	0.30	0.26	0.27
	ItalyPowerDemand	**0.79**	**0.81**	0.65	0.68	0.62	0.71	0.74	0.71	0.50	0.69	0.61
	FISH	**0.55**	**0.53**	0.38	0.38	0.26	0.47	0.52	0.44	0.14	0.49	0.39
	SwedishLeaf	**0.64**	0.62	0.47	0.48	0.31	0.45	0.51	0.43	0.06	**0.66**	0.24
	ToeSegmentation1	0.58	0.65	**0.76**	**0.72**	0.48	0.59	0.68	0.58	0.53	0.65	0.48
	ToeSegmentation2	0.62	0.75	0.62	0.59	0.54	0.64	0.75	0.64	**0.82**	**0.82**	0.78
	Average Prefix	**0.64**	**0.64**	0.56	0.58	0.47	0.59	0.61	0.59	0.39	0.55	0.49
Subsequence Dataset	Coffee	0.54	0.57	0.54	**0.71**	0.46	**0.71**	0.57	**0.71**	0.54	0.50	0.64
	ECGFiveDays	0.57	0.56	0.57	0.57	0.52	**0.59**	0.57	0.58	0.50	**0.58**	0.53
	DistalPhalanxTW	0.51	0.56	0.55	**0.58**	0.40	**0.61**	0.45	0.48	0.30	0.51	0.31
	BirdChicken	0.40	0.55	0.50	0.60	0.35	**0.70**	**0.75**	**0.70**	0.50	0.50	0.60
	GunPoint	0.65	0.66	0.63	0.64	0.56	0.65	**0.71**	0.67	0.49	**0.77**	0.59
	CBF	0.71	0.71	0.65	0.70	0.48	**0.75**	0.74	**0.75**	0.33	0.55	0.36
	DPOC	0.60	**0.63**	0.62	0.62	**0.62**	0.62	0.62	0.59	0.58	0.58	0.60
	OliveOil	**0.50**	0.43	**0.53**	0.40	0.43	0.50	0.50	0.37	0.17	0.17	0.37
	Symbols	**0.52**	0.50	0.49	0.48	0.37	0.48	0.51	0.48	0.16	**0.68**	0.30
	FaceFour	**0.51**	0.49	**0.52**	0.40	0.39	0.35	0.38	0.35	0.30	0.40	0.26
	ItalyPowerDemand	0.66	**0.71**	0.61	0.63	0.60	0.70	**0.71**	0.70	0.50	0.68	0.52
	FISH	0.35	**0.40**	0.31	0.31	0.17	0.31	0.30	0.27	0.14	**0.57**	0.21
	SwedishLeaf	0.42	**0.44**	0.39	0.40	0.20	0.33	0.38	0.31	0.06	**0.60**	0.11
	ToeSegmentation1	0.67	0.68	0.66	0.67	0.55	0.67	**0.69**	0.66	0.53	**0.69**	0.48
	ToeSegmentation2	0.71	0.78	0.70	0.68	0.58	0.69	0.79	0.69	**0.82**	**0.82**	0.73
	Average Subsequence	0.55	**0.58**	0.55	0.56	0.45	0.58	**0.58**	0.56	0.39	0.57	0.44
Suffix Dataset	Coffee	0.54	0.57	0.61	0.61	0.54	**0.68**	**0.68**	**0.71**	0.54	0.66	0.57
	ECGFiveDays	0.57	0.61	0.55	0.58	0.52	**0.63**	0.62	**0.62**	0.50	0.50	0.53
	DistalPhalanxTW	0.56	0.58	0.53	0.55	0.50	**0.61**	**0.58**	0.57	0.30	0.30	0.28
	BirdChicken	0.45	0.55	**0.70**	**0.60**	0.40	0.45	0.45	0.45	0.50	0.50	0.50
	GunPoint	0.61	**0.65**	0.57	0.60	0.65	0.65	**0.66**	0.62	0.49	0.49	0.58
	CBF	0.61	0.72	0.71	0.73	0.42	**0.78**	0.78	**0.78**	0.33	0.33	0.34
	DPOC	0.68	**0.71**	0.67	0.68	0.59	0.70	**0.71**	0.68	0.58	0.58	0.55
	OliveOil	0.17	**0.30**	0.23	**0.30**	0.27	0.17	0.20	0.13	0.17	0.13	0.17
	Symbols	**0.55**	0.52	0.51	0.53	0.38	0.51	**0.55**	0.50	0.16	0.53	0.24
	FaceFour	0.38	**0.45**	0.27	0.28	0.24	0.35	**0.39**	0.35	0.30	**0.39**	0.17
	ItalyPowerDemand	0.80	0.81	**0.82**	0.77	0.69	0.82	**0.85**	0.82	0.50	0.77	0.59
	FISH	0.36	0.37	0.35	0.35	0.16	0.37	**0.38**	0.34	0.14	**0.45**	0.14
	SwedishLeaf	0.53	0.54	0.53	0.53	0.30	0.49	**0.57**	0.48	0.06	**0.68**	0.18
	ToeSegmentation1	0.59	**0.64**	0.60	0.51	0.52	0.60	0.62	0.60	0.53	**0.82**	0.52
	ToeSegmentation2	0.71	**0.77**	0.62	0.62	0.55	0.68	0.73	0.69	**0.82**	0.69	**0.77**
	Average Suffix	0.54	**0.59**	0.55	0.55	0.45	0.57	**0.58**	0.56	0.39	0.52	0.41
	Average Total	0.58	**0.60**	0.55	0.56	0.46	0.58	**0.59**	0.57	0.39	0.55	0.45