PT$L^P$: Partial Transport $L^P$ 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) technique and its numerous extensions, as well as more recent methods based on optimal transport principles.

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. 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, the concept of barycenters, and parameterized geodesics. 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, optimal partial transport (OPT) problem, Kantorovich-Rubinstein norm and unnormalized optimal transport are some of the variants that fall under the category of "unbalanced optimal transport". 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 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 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. 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. Later, Zhang et al. utilized a similar approach to TL$^p$ while adding entropy regularization and introduced Time Adaptive OT (TAOT). Lastly, in Spatio-Temporal Alignments, Janati et al. 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 $\mu$ and $\nu$ in $\mathcal{P}(\Omega)$, and a lower semi-continuous cost function $c : \Omega^2 \to \mathbb{R}^+$ by:

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

(1)

Here, $\Pi(\mu, \nu)$ is the set of all joint probability measures whose marginals are $\mu$ and $\nu$. We represent this by $\pi_1 \# \gamma = \mu$ and $\pi_2 \# \gamma = \nu$, where $\pi_1$ and $\pi_2$ 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 $\Omega$. Let $\lambda \geq 0$ denote the penalty for mass creation or destruction. Then the OPT problem is defined as:

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

(2)

where

$$\Pi_c(\mu, \nu) := \{ \gamma \in \mathcal{M}_+(\Omega^2) : \pi_1 \# \gamma \leq \mu, \pi_2 \# \gamma \leq \nu \},$$

$\pi_1 \# \gamma \leq \mu$ indicates that $\pi_1 \# \gamma$ is dominated by $\mu$, 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 $\|\mu\|_{\text{TV}}$ is the total variation (and the total mass) of $\mu$, analogously for $\|\nu\|_{\text{TV}}$. When the transportation cost $c(x, y)$ is a metric, $\text{OPT}_{\lambda, c}(\cdot, \cdot)$ 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 OT 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 [52] 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 $(\theta, \cdot)_{\#} \mu$ where $(\cdot, \cdot) : \Omega^2 \to \mathbb{R}$ denotes the inner product. Then for $\mu, \nu \in \mathcal{P}_p(\Omega)$ we can define the Sliced-OT (SOT) as:

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

(3)

where $\sigma \in \mathcal{P}(\mathbb{S}^{d-1})$ is a probability measure such that $\text{supp}(\sigma) = \mathbb{S}^{d-1}$, 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:

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

(4)

where $\lambda \in L^1(\sigma; \mathbb{R}^+)$ 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, 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 $\mu$ and $\nu$ 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
We name it as the transport \( TL \). To do so, we first expand the definition of transport, rather than probability measures. Specifically, we define a signal as the pair \((f, \mu)\) for any \( f \in L^\infty(\Omega) \) and \( \mu \in P(\Omega; \mathbb{R}^k) \) for every \( k \)-dimensional signal, denoting, for instance, the impulse train used for sampling the signal, and \( \mu \) is a metric space. Intuitively, the \( TL \) distance measures the OT between measures \( \mu \) and \( \nu \), respectively. On the right, the optimal transport plan is visualized, accompanied by the corresponding transportation cost.

**Transport \( L^p \) Distances.** Following [7], a multi-channel signal with \( k \) channels can be defined as the pair \((f, \mu)\) for \( \mu \in P_p(\Omega) \) 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 \[ Q_p(\Omega; \mathbb{R}^k) := \{ (f, \mu) | \mu \in P_p(\Omega), f \in L^p(\mu; \mathbb{R}^k) \} \]

We name it as the transport \( L^p \) space. The \( TL^p_\beta \) distance between two such \( k \)-dimensional signals \((f, \mu)\) and \((g, \nu)\) in \( Q_p(\Omega; \mathbb{R}^k) \) is defined as:

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

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

\[
\lim_{\beta \to 0} TL^p_\beta((f, \mu), (g, \nu)) = \| f - g \|_{L^\infty(\mu)} \quad \text{if } \mu = \nu
\]

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

Hence, the \( TL^p_\beta \) 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 \((f, \mu)\) where \( \mu \in M_+(\Omega) \)
Furthermore, for the empirical distributions, we refer to Section C in the appendix for the proofs of the above theorems and a detailed discussion.

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

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

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:

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

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 $\gamma$ as the optimal $\gamma$ will be induced by a 1-1 mapping.

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

**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 (8), there exists a minimizer $\gamma \in \Pi_{\leq}(1_M, 1_N)$ that is induced by a 1-1 mapping. That is, the optimal $\gamma$ satisfies $\gamma_{ij} \in \{0, 1\}$ for each $i, j$, and each row and column of $\gamma$ contains at most one nonzero element.

**Theorem 3.2.** $(Q_+(\Omega; \mathbb{R}^k), PTL^p_{\beta,\lambda})$ 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 $Q^p_+(\Omega; \mathbb{R}^k)$.

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

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

$$\lim_{\beta \to \infty} PTL^p_{\beta,\lambda}((f, \mu), (g, \nu)) = OPT_{\lambda}(f \# \mu, g \# \nu),$$
where \( \mu \wedge \nu \) is the minimum of measure \( \mu, \nu \),

\[
\|f - g\|_{\text{L}^p(\mu \wedge \nu), 2\lambda}^p := \int_{\Omega} \|f - g\|^p \wedge 2\lambda d(\mu \wedge \nu).
\]
and \( \|\mu - \nu\|_{\text{TV}} \) 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 \( \gamma \) of all zeros, i.e., \( \text{PTL}^p_{\beta, \theta}(\mu, f), (\nu, g)) = 0 \) for all \((\mu, f), (\nu, g) \in Q^p_+(\Omega; \mathbb{R}^k) \).

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

\[
\text{TL}^p_{\beta}(f, \mu), (g, \nu)) = \text{OT}(\hat{\mu}, \hat{\nu})
\]

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

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

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 PTLP and OPT (see proposition \([3]\)), we can define Sliced PTLP as

\[
\text{SPTL}^p_{\beta, \lambda}(f, \mu), (g, \nu)) = \int_{\mathbb{S}^{d+k-1}} \text{OPT}(\lambda(\theta_\# \hat{\mu}, \theta_\# \hat{\nu})d\sigma(\theta)
\]

where \( \lambda \) can be defined as an \( L^1(\sigma, \mathbb{R}^{d+k}) \) function of \( \theta \). Note that STL\( ^p_{\beta} \) and SPTL\( ^p_{\beta, \lambda} \) are metrics on \( Q(\Omega; \mathbb{R}^k) \) and \( Q_+(\Omega; \mathbb{R}^k) \), 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 PTLP distance on different classes of signals in this experiment. We generate the following two classes of signals on the domain \([0, 1]\): \( S_0 \) is the class of signals with translations of one positive Gaussian bump, whereas \( S_1 \) 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 \( L^p \), TLP and PTLP distance matrices. It can be seen that PTLP 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\( P^p \) approach tends to misinterpret the noise as the predominant pattern.

#### 4.2 Nearest Neighbor Classification

**Experiment setup**

To demonstrate the effectiveness of our proposed PTLP metric and its sliced variant SPTLP, we
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 \( \beta \) and \( \lambda \)**

To find the optimal \( \beta \) and \( \lambda \) for \( SPTL_{\beta,\lambda}^P \), 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 \( \beta \) is set to be \( \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 10^3, 10^4\} \), and \( \lambda \) 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 \( SPTL_{\beta,\lambda}^P \), we also need to specify the slices, i.e. \( \theta \)'s for 1 dimensional projections. We obtain the optimal \( \beta \) from \( PTL_{\beta,\lambda}^P \). As the amount of mass that should be transported may vary across slices, we adopt the strategy to search for the best \( \lambda \) for the most informative slice, and then set \( \lambda \)'s accordingly for other slices. We set \( \theta_0 \) to be the first principle component of all signals. Note that \( \theta_0 \) vanishes at dimensions corresponding to \( x^{\beta - \frac{1}{p}} \), but concentrates on \( f(x) \) in \( T_{\beta,f,p}(x) = [x^{\beta - \frac{1}{p}}; f(x)] \) (refer to Eq. 12 and Eq. 13). Similarly, we implement grid search for best \( \lambda_{\theta_0} \) corresponding to \( \theta_0 \). Given \( \theta_0 \) and \( \lambda_{\theta_0} \), for a specific slice \( \theta \), \( \lambda_{\theta} = \langle \theta, \theta_0 \rangle \lambda_{\theta_0} \), 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, \( SPTL_{\theta}^P \) 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 \( SPTL_{\theta}^P \) and \( STL_{\theta}^P \) 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 STLP and STLp provide significantly faster solutions when accelerated by parallel computation with respect to slices.

<table>
<thead>
<tr>
<th>Method</th>
<th>PTLp</th>
<th>STLP</th>
<th>TLp</th>
<th>STLp</th>
<th>OT</th>
<th>SDTW (\gamma = 0.01)</th>
<th>SDTW (\gamma = 1)</th>
<th>DTW</th>
<th>WDTW</th>
<th>DDTW</th>
<th>Lp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adiac</td>
<td>0.23</td>
<td>0.26</td>
<td>0.23</td>
<td>0.22</td>
<td>0.08</td>
<td>0.21</td>
<td>0.19</td>
<td>0.19</td>
<td>0.03</td>
<td>0.29</td>
<td>0.25</td>
</tr>
<tr>
<td>ArrowHead</td>
<td>0.49</td>
<td>0.51</td>
<td>0.44</td>
<td>0.41</td>
<td>0.45</td>
<td>0.45</td>
<td>0.43</td>
<td>0.39</td>
<td>0.52</td>
<td>0.55</td>
<td>0.25</td>
</tr>
<tr>
<td>BeetleFly</td>
<td>0.65</td>
<td>0.80</td>
<td>0.55</td>
<td>0.55</td>
<td>0.40</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.50</td>
<td>0.75</td>
<td>0.55</td>
</tr>
<tr>
<td>DSR</td>
<td>0.68</td>
<td>0.53</td>
<td>0.45</td>
<td>0.47</td>
<td>0.30</td>
<td>0.38</td>
<td>0.37</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.60</td>
</tr>
<tr>
<td>DPOAG</td>
<td>0.68</td>
<td>0.68</td>
<td>0.65</td>
<td>0.64</td>
<td>0.59</td>
<td>0.65</td>
<td>0.63</td>
<td>0.65</td>
<td>0.42</td>
<td>0.42</td>
<td>0.55</td>
</tr>
<tr>
<td>ECG200</td>
<td>0.78</td>
<td>0.74</td>
<td>0.78</td>
<td>0.72</td>
<td>0.64</td>
<td>0.74</td>
<td>0.77</td>
<td>0.74</td>
<td>0.36</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>Ham</td>
<td>0.52</td>
<td>0.57</td>
<td>0.51</td>
<td>0.50</td>
<td>0.54</td>
<td>0.55</td>
<td>0.56</td>
<td>0.57</td>
<td>0.49</td>
<td>0.49</td>
<td>0.45</td>
</tr>
<tr>
<td>Herring</td>
<td>0.63</td>
<td>0.69</td>
<td>0.61</td>
<td>0.61</td>
<td>0.47</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
<td>0.59</td>
<td>0.59</td>
<td>0.56</td>
</tr>
<tr>
<td>Lightning?</td>
<td>0.49</td>
<td>0.44</td>
<td>0.40</td>
<td>0.49</td>
<td>0.38</td>
<td>0.55</td>
<td>0.51</td>
<td>0.53</td>
<td>0.26</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td>OSUleaf</td>
<td>0.45</td>
<td>0.41</td>
<td>0.40</td>
<td>0.33</td>
<td>0.32</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.10</td>
<td>0.52</td>
<td>0.19</td>
</tr>
<tr>
<td>Plane</td>
<td>0.80</td>
<td>0.80</td>
<td>0.69</td>
<td>0.68</td>
<td>0.48</td>
<td>0.68</td>
<td>0.72</td>
<td>0.67</td>
<td>0.14</td>
<td>0.14</td>
<td>0.50</td>
</tr>
<tr>
<td>ShapeletSim</td>
<td>0.52</td>
<td>0.49</td>
<td>0.48</td>
<td>0.53</td>
<td>0.50</td>
<td>0.51</td>
<td>0.57</td>
<td>0.51</td>
<td>0.50</td>
<td>0.50</td>
<td>0.49</td>
</tr>
<tr>
<td>SyntheticControl</td>
<td>0.84</td>
<td>0.81</td>
<td>0.76</td>
<td>0.83</td>
<td>0.55</td>
<td>0.87</td>
<td>0.84</td>
<td>0.87</td>
<td>0.17</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td>Trace</td>
<td>0.75</td>
<td>0.72</td>
<td>0.79</td>
<td>0.75</td>
<td>0.44</td>
<td>0.81</td>
<td>0.78</td>
<td>0.82</td>
<td>0.23</td>
<td>0.24</td>
<td>0.61</td>
</tr>
<tr>
<td>Wine</td>
<td>0.50</td>
<td>0.59</td>
<td>0.46</td>
<td>0.54</td>
<td>0.59</td>
<td>0.44</td>
<td>0.46</td>
<td>0.46</td>
<td>0.50</td>
<td>0.56</td>
<td>0.50</td>
</tr>
<tr>
<td>Average Prefix</td>
<td>0.59</td>
<td>0.60</td>
<td>0.55</td>
<td>0.55</td>
<td>0.45</td>
<td>0.54</td>
<td>0.55</td>
<td>0.54</td>
<td>0.33</td>
<td>0.45</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>PTLp</th>
<th>STLP</th>
<th>TLp</th>
<th>STLp</th>
<th>AveragePrefix</th>
<th>AverageSubsequence</th>
<th>AverageSuffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adiac</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>ArrowHead</td>
<td>0.46</td>
<td>0.49</td>
<td>0.51</td>
<td>0.47</td>
<td>0.46</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>BeetleFly</td>
<td>0.55</td>
<td>0.50</td>
<td>0.65</td>
<td>0.55</td>
<td>0.60</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>DSR</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>DPOAG</td>
<td>0.63</td>
<td>0.65</td>
<td>0.65</td>
<td>0.63</td>
<td>0.54</td>
<td>0.60</td>
<td>0.58</td>
</tr>
<tr>
<td>ECG200</td>
<td>0.71</td>
<td>0.69</td>
<td>0.59</td>
<td>0.62</td>
<td>0.53</td>
<td>0.62</td>
<td>0.67</td>
</tr>
<tr>
<td>Ham</td>
<td>0.54</td>
<td>0.49</td>
<td>0.49</td>
<td>0.52</td>
<td>0.49</td>
<td>0.52</td>
<td>0.49</td>
</tr>
<tr>
<td>Herring</td>
<td>0.45</td>
<td>0.56</td>
<td>0.56</td>
<td>0.58</td>
<td>0.48</td>
<td>0.50</td>
<td>0.53</td>
</tr>
<tr>
<td>Lightning?</td>
<td>0.38</td>
<td>0.38</td>
<td>0.37</td>
<td>0.37</td>
<td>0.29</td>
<td>0.37</td>
<td>0.45</td>
</tr>
<tr>
<td>OSUleaf</td>
<td>0.42</td>
<td>0.51</td>
<td>0.51</td>
<td>0.31</td>
<td>0.24</td>
<td>0.34</td>
<td>0.41</td>
</tr>
<tr>
<td>Plane</td>
<td>0.57</td>
<td>0.51</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.51</td>
<td>0.58</td>
</tr>
<tr>
<td>ShapeletSim</td>
<td>0.49</td>
<td>0.46</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td>0.51</td>
<td>0.53</td>
</tr>
<tr>
<td>SyntheticControl</td>
<td>0.69</td>
<td>0.63</td>
<td>0.63</td>
<td>0.71</td>
<td>0.76</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td>Trace</td>
<td>0.63</td>
<td>0.56</td>
<td>0.67</td>
<td>0.68</td>
<td>0.36</td>
<td>0.78</td>
<td>0.64</td>
</tr>
<tr>
<td>Wine</td>
<td>0.56</td>
<td>0.67</td>
<td>0.63</td>
<td>0.59</td>
<td>0.54</td>
<td>0.29</td>
<td>0.61</td>
</tr>
<tr>
<td>AveragePrefix</td>
<td>0.50</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
<td>0.41</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>AverageSubsequence</td>
<td>0.50</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
<td>0.41</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>AverageSuffix</td>
<td>0.50</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
<td>0.41</td>
<td>0.49</td>
<td>0.53</td>
</tr>
</tbody>
</table>

4.3 Computation efficiency using Sliced PTLp

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 [islearn], which are accelerated by [numba]. TLp and PTLp 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. STLP, SPTLP can be accelerated by [numba]. For STLP and SPTLP, we set the number of projections to be 50. Note, the computation of STLP and SPTLP 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Worst-case Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTL$^p$</td>
<td>$O(N^3 + N^2(d + k))$</td>
</tr>
<tr>
<td>SPTL$^p$</td>
<td>$O(LN(\log(N) + (d + k)))$</td>
</tr>
<tr>
<td>TL$^p$</td>
<td>$O(N^3 + N^2(d + k))$</td>
</tr>
<tr>
<td>STL$^p$</td>
<td>$O(LN(\log(N) + (d + k)))$</td>
</tr>
<tr>
<td>OT</td>
<td>$O(k(N^3 + N^2d))$</td>
</tr>
<tr>
<td>*DTW</td>
<td>$O(N^2k)$</td>
</tr>
<tr>
<td>$L^p$</td>
<td>$O(Nk)$</td>
</tr>
</tbody>
</table>

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 PTL$^p_{\beta,\lambda}$ as $\beta$ 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


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 $k \geq 1$.
- $\Omega$: 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 $\Omega, \Omega^2$ respectively.
- $\mathcal{M}_+(\Omega), \mathcal{M}_+(\Omega^2)$: the set of positive Radon measures on $\Omega, \Omega^2$ respectively.
- $\text{spt}(\mu)$ for $\mu \in \mathcal{M}(\Omega)$: the support of the measure $\mu$.
- $\|\mu\|_{TV}$ where $\mu \in \mathcal{M}(\Omega)$: the total variation of $\mu$. If $\mu$ is positive, $\|\mu\|_{TV} = \mu(\Omega)$.
- 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

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

then we say $\gamma_n$ converges weakly in measure to $\gamma$, 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 $\gamma_n \rightharpoonup \gamma$.

- $\pi_1, \pi_2$: the canonical projections on $\Omega^2$, i.e. $\pi_1((x, y)) = x, \pi_2((x, y)) = y$.

- $f_\# \gamma$: the push-forward of the measure $\gamma$ by $f$, i.e. $f_\# \gamma(A) = \gamma(f^{-1}(A))$ for any Borel set $A$.

- $\mu \wedge \nu$: minimum measure between $\mu$ and $\nu$. 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 $\mu, \nu$ 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 $\mu, \nu$, 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 $\mu(\Omega) = \nu(\Omega)$.

- $\Pi_{<}(\mu, \nu)$ where $\mu, \nu \in \mathcal{M}_+(\Omega)$: the set of Kantorovich plans between $\mu, \nu$ for the OPT problem, i.e. $\Pi_{<}(\mu, \nu) := \{ \gamma \in \mathcal{M}_+(\Omega^2) : (\pi_1)_\# \gamma \leq \mu, (\pi_2)_\# \gamma \leq \nu \}$.

- $(C^0_b(\Omega^2), \| \cdot \|_{sup})$: the space of continuous and bounded functions on $\Omega^2$. Note, the dual space of $C^0_b(\Omega^2)$ is the space of Radon measures $\mathcal{M}(\Omega^2)$.

- $\| \cdot \|_{L^p(\mu)}$, $\| \cdot \|_{L^p(\mu), 2\lambda}$ where $\mu \in \mathcal{M}_+(\Omega)$: the $L^p$ norm and truncated $L^p$ norm with respect to reference measure $\mu$, i.e.

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

$$\|f\|_{L^p(\mu), 2\lambda} = \left( \int_{\Omega} |f(x)|^p \land 2\lambda \, d\mu(x) \right)^{1/p}.$$
We will also make use of the closure, in the weak topology, of the space \( \Pi \).

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

**Proof.**

Let \( \{ \mu_n \} \) be a sequence in \( \Pi \). We need to show that if \( \mu_n \rightharpoonup \mu \) in the weak topology, then \( \mu_n \rightharpoonup \mu \) in the weak topology as well. This is equivalent to showing that for any continuous function \( f \),

\[
\int f \, d\mu_n \to \int f \, d\mu \quad \text{as} \quad n \to \infty.
\]

For any \( \epsilon > 0 \), there exists \( N \) such that for all \( n \geq N \),

\[
|\int f \, d\mu_n - \int f \, d\mu| < \epsilon.
\]

This completes the proof.

**Proposition B.2.**

Given that weak convergence in measure preserves inequalities,

\[
\Pi(\Omega) = \Pi(\Omega; \mathbb{R}^k) = \text{the closure, in the weak topology, of the space } \Pi(\Omega)
\]

**Proposition B.1.**

Given a sequence \( \{ \mu_n \} \subseteq \mathcal{M}_+(\Omega) \), 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 \( f \in C^0_b(\Omega) \), we have

\[
\int f \, d\mu_0 = \lim_{n \to \infty} \int f \, d\mu_n \leq \int f \, d\mu,
\]

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 \( \Pi(\mu, \nu) \), which follows similarly to the analogous result for the space \( \Pi(\mu, \nu) \). A similar result has been proved when \( \mu, \nu \) are continuous measure, for example [18, Lemma 2.2]. For completeness we include a proof in the general case.

**Proposition B.2.**

The set \( \Pi(\mu, \nu) \) is sequentially compact in the weak topology.

**Proof.**

Pick a sequence \( \{ \gamma_n \} \subseteq \Pi(\mu, \nu) \). We have \( \gamma_n \) is bounded with respect to total variation. Indeed,

\[
\|\gamma_n\|_{TV} = \gamma_n(\Omega) \leq \mu(\Omega), \quad \forall n
\]

In addition, we will show \( \gamma_n \) is tight. Pick \( \epsilon > 0 \), since \( \mu, \nu \) are inner regular, there exists compact set \( K \subseteq \Omega \) 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 \((\gamma_n)\) 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 \(\pi_{1\#} \gamma_n \rightharpoonup \pi_{1\#} \gamma\) and \(\pi_{2\#} \gamma_n \rightharpoonup \pi_{2\#} \gamma\). Indeed, pick \(f \in C_0^0(\Omega)\), \(f\) can be regarded as a function in \(C_0^0(\Omega^2)\) whose value is independent to the second input \(y\). Thus, we have

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

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

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

The identity in \(\mathcal{Q}_P^+(\Omega; \mathbb{R}^k)\) is defined as: \((f, \mu) = (g, \nu)\) if and only if \(\mu = \nu\) and \(f = g\), \(\mu\)-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 \((f, \mu) \mapsto (\text{id} \times f)_{\#} \mu\) 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 \((\text{id} \times f)_{\#} \mu\).

**Proposition C.2.** The mapping \(T : \mathcal{Q}_P^+(\Omega; \mathbb{R}^k) \to \text{Ran}(T) \subset \mathcal{M}_+(\Omega \times \mathbb{R}^k)\) 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

\[
T((f, \mu))(\{(x, f(x)) : x \in A\}) = (\text{id} \times f)_{\#} \mu(\{(x, f(x)) : x \in A\})
\]

\[
= \mu(A)
\]

\[
> \nu(A)
\]

\[
= (\text{id} \times g)_{\#} \nu(\{(x, g(x)) : x \in A\})
\]

\[
\geq (\text{id} \times g)_{\#} \nu(\{(x, f(x)) : x \in A, f(x) = g(x)\})
\]

\[
= (\text{id} \times g)_{\#} \nu(\{(x, f(x)) : x \in A\})
\]

\[
= T((g, \nu))(\{(x, f(x)) : x \in A\})
\]

where (15) follows from the fact \((\text{id} \times g)_{\#} \nu\) 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

\[
(\text{id} \times f)_{\#} \mu(\{(x, f(x)) : x \in B\}) = \mu(B) > 0
\]

and

\[
(\text{id} \times f)_{\#} \nu(\{(x, f(x)) : x \in B\}) = \nu\{x : f(x) = g(x), x \in B\}
\]

\[
= \mu\{x : f(x) = g(x), x \in B\}
\]

\[
= 0.
\]

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

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

It is straightforward to show $D$ is metric and is equivalent to the $\ell^p$ 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 [53 Theorem 2.2], where $\hat{\mu}, \hat{\nu} \in \mathcal{M}_+(\Omega \times \mathbb{R}^k)$:

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

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

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

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

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

Therefore,

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

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

$$\hat{\gamma}_1(A \times \mathbb{R}^k \times B \times \mathbb{R}^k) = ((id \times f), (id \times g))\# \gamma_1(A \times \mathbb{R}^k \times B \times \mathbb{R}^k)$$

$$= \gamma_1 \left( \{(x, y) : (x, f(x), y, g(y)) \in A \times \mathbb{R}^k \times B \times \mathbb{R}^k \} \right)$$

$$= \gamma_1(A \times B)$$

$$> \gamma_2(A \times B)$$

$$= \gamma_2 \left( \{(x, y) : (x, f(x), y, g(y)) \in A \times \mathbb{R}^k \times B \times \mathbb{R}^k \} \right)$$

$$= ((id \times f), (id \times g))\# \gamma_2(A \times \mathbb{R}^k \times B \times \mathbb{R}^k)$$

$$= \hat{\gamma}_2(A \times \mathbb{R}^k \times B \times \mathbb{R}^k).$$

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

To show surjectivity, take any $\hat{\gamma} \in \Pi_{\leq}(\hat{\mu}, \hat{\nu})$. Take any $\hat{A}, \hat{B}$ such that $0 = (\hat{\mu} \times \hat{\nu})(\hat{A} \times \hat{B}) = \mu(A)\nu(B)$. 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 $\text{spt}(\hat{\gamma}) \subseteq \text{spt}(\hat{\mu} \times \hat{\nu})$. It follows in a similar way that $\text{spt}(F(\gamma)) \subseteq \text{spt}(\hat{\mu} \times \hat{\nu})$ 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 $\hat{A} \in \text{spt}(\hat{\mu})$ and $\hat{B} \in \text{spt}(\hat{\nu})$. In particular, we may assume that $A = \{(x, f(x)) : x \in A\}$ and $B = \{(y, g(y)) : y \in B\}$ for some $A, B \subseteq \Omega$. Now

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

Thus $F\gamma = \hat{\gamma}$ and so $F$ is surjective.
For each $\gamma \in \Pi(\mu, \nu)$ we continue to let $\hat{\gamma} = F(\gamma)$ denote the corresponding measure in $\Pi(\hat{\mu}, \hat{\nu})$. We have

$$\hat{C}(\hat{\gamma}; \hat{\mu}, \hat{\nu}, D^p_\beta, \lambda) = \int_{(\Omega \times \mathbb{R}^k)^2} D^p_\beta((x, \hat{x}), (y, \hat{y})) \, d\hat{\gamma}(x, \hat{x}, y, \hat{y}) + \lambda(\|\hat{\mu}\|_{TV} + \|\hat{\nu}\|_{TV} - 2\|\hat{\gamma}\|_{TV})$$

$$= \int_{\Omega^2} \frac{1}{\beta}\|x - y\|^p + \|f(x) - g(y)\|^p \, d\gamma(x, y) + \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma\|_{TV})$$

$$= C(\gamma; (f, \mu), (g, \nu), \beta, \lambda)$$

(17)

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

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

which completes the proof. \hfill \Box

**Remark C.4.** Proposition C.2 and C.3 imply that $(Q^p_\beta(\Omega), \text{PTL}^p_{\beta, \lambda})$ 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 $(\Omega \times \mathbb{R}^k) \cup \{\infty\}$ where $\infty$ 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, $\Pi(\mu, \nu)$ is compact in the weak topology by Proposition B.2 and $\gamma \mapsto \hat{C}(\gamma; \hat{\mu}, \hat{\nu}, D^p_\beta, \lambda)$ 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. \hfill \Box

### D The PTL$^p$ Problem for Extreme $\beta$

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

$$\text{PTL}_{\beta, \lambda}^p((f, \mu), (g, \nu)) = \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\Omega^2} \left(\frac{1}{\beta}\|x - y\|^p + \|f(x) - g(y)\|^p\right) + 2\lambda \, d\gamma(x, y)$$

$$+ \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma\|_{TV})$$

(18)

**Proof.** Similar to the last section, the PTL$^p$ problem can be written as the following OPT problem between $\mu, \nu$ defined as follows:

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

(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 $\gamma$ such that $\gamma(S) = 0$ on $S = \{(x, y) : c(x, y) \geq 2\lambda\}$ and we complete the proof. \hfill \Box
Now we discuss the extreme cases $\beta = 0$ and $\beta = \infty$.

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

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

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

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

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

Since

$$\frac{1}{\beta_n} \left( \int_{\Omega^2} \|x - y\|^p \land 2\lambda d\gamma(x, y) + \beta_n \lambda \left( \|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma\|_{TV} \right) \right)$$

then

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

So $\text{OPT}_{\|\cdot\|_p \land 2\lambda}(\mu, \nu) \leq O(\beta_n) \to 0$.

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

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

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

$$\text{OPT}_{\|\cdot\|_p \land 2\lambda}(\mu, \nu) = \int_{\Omega^2} \|x - y\|_p \land 2\lambda d\gamma_0(x, y).$$

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

$$\text{PTL}^p_{\mu_0,\lambda}(f, \mu, (g, \nu)) := \inf_{\mu_0 \leq \mu \leq \nu} \int_{\Omega} \|f(x) - g(x)\|^p \land 2\lambda d\mu_0(x) + \lambda \left( \|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\mu_0\|_{TV} \right). \quad (21)$$

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 $\mu_0$, we have

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

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

$$\text{PTL}^p_{\mu_0,\lambda}(f, \mu, (g, \nu)) = \|f - g\|_{\mu \wedge \nu, 2\lambda} + \lambda(\|\mu - \nu\|_{TV}). \quad (22)$$

18
Since \( C^0_0(\Omega) \) is dense in \( L^p(\mu), L^p(\nu) \), for any \( \varepsilon > 0 \) there exists \( f_\varepsilon, g_\varepsilon \in C^0_0(\Omega) \) such that
\[
\int_{\Omega^2} \|f(x) - f_\varepsilon(x)\|^p \, d\mu(x) \leq \varepsilon, \quad \int_{\Omega^2} \|g(y) - g_\varepsilon(y)\|^p \, d\nu(y) \leq \varepsilon.
\]

Now,
\[
\left( \int_{\Omega^2} \|f(x) - g(y)\|^p \, d\gamma_n(x, y) \right)^{\frac{1}{p}} \geq \left( \int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \, d\gamma_n(x, y) \right)^{\frac{1}{p}}
\]
\[
- \left( \int_{\Omega^2} \|f(x) - f_\varepsilon(x)\|^p \, d\gamma_n(x, y) \right)^{\frac{1}{p}}
\]
\[
- \left( \int_{\Omega^2} \|g_\varepsilon(y) - g(y)\|^p \, d\gamma_n(x, y) \right)^{\frac{1}{p}}
\]
\[
\geq \left( \int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \, d\gamma_n(x, y) \right)^{\frac{1}{p}} - 2\varepsilon
\]
\[
\Rightarrow \left( \int_{\Omega^2} \|f(x) - g(y)\|^p \, d\gamma_0(x, y) \right)^{\frac{1}{p}} - 2\varepsilon.
\]
So,
\[
\liminf_{n \to \infty} \text{PTL}^p_{\beta_n, \lambda}(\{(f, \mu), (g, \nu)\}) \geq \int_{\Omega^2} \|f_\varepsilon(x) - g_\varepsilon(y)\|^p \, d\gamma_0(x, y) - K\varepsilon
\]
\[
+ \lambda \left( \|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma_0\|_{TV} \right)
\]
\[
\geq \int_{\Omega^2} \|f(x) - g(y)\|^p \, d\gamma_0(x, y) - K'\varepsilon
\]
\[
+ \lambda \left( \|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma_0\|_{TV} \right)
\]
\[
\geq \text{PTL}^p_{0, \lambda}(\{(f, \mu), (g, \nu)\}) - K'\varepsilon
\]
for constants \( K, K' \). Taking \( \varepsilon \to 0 \) completes the proof. \( \square \)

Based on the above theorem, we can extend the \( \text{PTL}^p \) distance for \( \beta = 0 \).

**Corollary D.3.** When \( \mu = \nu \), we have
\[
\text{PTL}^p_{0, \lambda}(\{(f, \mu), (g, \mu)\}) = \|f - g\|^p_{L^p(\mu), 2\lambda}.
\]

The proof is straightforward.

Next, we discuss the case \( \beta = \infty \).

**Theorem D.4.** For any positive sequence \( \beta_n \to \infty \), we have
\[
\lim_{n \to \infty} \text{PTL}^p_{\beta_n, \lambda}(\{(f, \mu), (g, \nu)\}) = \text{OPT}_\lambda(f_\#, \mu, g_\#, \nu).
\] (23)

**Proof.** Without loss of generality, we can assume \( \beta_n \) is a monotonic increasing sequence. Note, it is straightforward to show (e.g. see Proposition C.3) by setting the \( \frac{1}{p} \) term to be 0:
\[
\text{OPT}_\lambda(f_\#, \mu, g_\#, \nu) = \text{OPT}_{c, \infty, f, \#, g, \#, \lambda}(\mu, \nu)
\]
\[
= \inf_{\gamma \in \Pi_{\infty, f, \#, g, \#, \lambda}(\mu, \nu)} \int_{\Omega^2} \|f(x) - g(y)\|^p \, d\gamma(x, y) + \lambda \left( \|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma\|_{TV} \right)
\] (24)
where \( c_{\infty, f, \#, g, \#, \lambda}(x, y) := \|f(x) - g(y)\|^p \). For each \( n \) and \( \gamma \), we have
\[
\int_{\Omega^2} \frac{1}{\beta_n} \|x - y\|^p + \|f(x) - g(y)\|^p \, d\gamma(x, y) + \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma\|_{TV})
\]
\[
\geq \int_{\Omega^2} \|f(x) - g(y)\|^p \, d\gamma(x, y) + \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\|\gamma\|_{TV}).
\] (25)
Taking the infimum on both sides and passing to the limit, we have
\[
\lim_{n \to \infty} \inf P TL_{\beta, \lambda_n}^p ((f, \mu), (g, \nu)) \geq OPT_{c_{f,g}, \lambda}(\mu, \nu) \tag{26}
\]
For the other direction, we let \(\gamma_\infty\) be an optimal transportation plan for \(OPT_{c_{f,g}, \lambda}(\mu, \nu)\), then
\[
\limsup_{n \to \infty} P TL_{\beta, \lambda_n}^p ((\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 d\gamma_\infty(x, y) + \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\gamma_\infty\|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 d\gamma_\infty(x, y) + \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\gamma_\infty\|TV\) \\
= \int_{\Omega^2} \|f(x) - g(y)\| \wedge 2\lambda d\gamma_\infty(x, y) + \lambda(\|\mu\|_{TV} + \|\nu\|_{TV} - 2\gamma_\infty\|TV\)
\]
where the third line follows from the Monotone convergence theorem (or Beppo Levi’s lemma). Thus
\[
\limsup_{n \to \infty} P TL_{\beta, \lambda_n}^p ((f, \mu), (g, \nu)) \leq OPT_{c_{\infty, f,g}, \lambda}(\mu, \nu) \tag{27}
\]
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.

<table>
<thead>
<tr>
<th>Prefix/Dataset</th>
<th>Subsequence/Dataset</th>
<th>Suffix/Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td><strong>PTL(^2)</strong></td>
<td><strong>SPTL(^2)</strong></td>
</tr>
<tr>
<td>Coffee</td>
<td>0.79</td>
<td>0.68</td>
</tr>
<tr>
<td>ECgFiveDays</td>
<td>0.58</td>
<td>0.60</td>
</tr>
<tr>
<td>DistalPhalanxTW</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>BirdChicken</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>GunPoint</td>
<td>0.63</td>
<td>0.71</td>
</tr>
<tr>
<td>CBF</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>DPOC</td>
<td>0.62</td>
<td>0.72</td>
</tr>
<tr>
<td>OliveOil</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Symbols</td>
<td>0.64</td>
<td>0.61</td>
</tr>
<tr>
<td>FaceFour</td>
<td>0.66</td>
<td>0.53</td>
</tr>
<tr>
<td>ItalyPowerDemand</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>FISH</td>
<td>0.55</td>
<td>0.53</td>
</tr>
<tr>
<td>SwedishLeaf</td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td>ToeSegmentation1</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>ToeSegmentation2</td>
<td>0.62</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Average Prefix</strong></td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Average Subsequence</strong></td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Average Suffix</strong></td>
<td>0.54</td>
<td>0.57</td>
</tr>
</tbody>
</table>