Two-Sample Test with Kernel Projected Wasserstein Distance

Jie Wang
Georgia Institute of Technology

Rui Gao
University of Texas at Austin

Yao Xie
Georgia Institute of Technology

Abstract

We develop a kernel projected Wasserstein distance for the two-sample test, an essential building block in statistics and machine learning: given two sets of samples, to determine whether they are from the same distribution. This method operates by finding the nonlinear mapping in the data space which maximizes the distance between projected distributions. In contrast to existing works about projected Wasserstein distance, the proposed method circumvents the curse of dimensionality more efficiently. We present practical algorithms for computing this distance function together with the non-asymptotic uncertainty quantification of empirical estimates. Numerical examples validate our theoretical results and demonstrate good performance of the proposed method.

1 INTRODUCTION

As a fundamental problem in statistical inference (Young et al., 2005), two-sample hypothesis testing aims to determine whether two sets of samples come from the same distribution or not. This problem has broad applications in scientific discovery fields. For example, it can be applied in anomaly detection (Chandola et al., 2009; Savage et al., 2014; Ahmed et al., 2016) to identify abnormal observations that follow a distinct distribution compared with typical observations. Similarly, in change-point detection (Poor and Hadjiliadis, 2008; Xie and Xie, 2021; Xie et al., 2021), two-sample testing is essential to detect abrupt changes in streaming data. Other notable examples include model criticism (Lloyd and Ghahramani, 2015; Chwialkowski et al., 2016; Binkowski et al., 2018), causal inference (Lopez-Paz and Oquab, 2018), and health care (Schober and Vetter, 2019).

Parametric or low-dimensional testing scenarios have been the main focus in classical literature. When extra knowledge about the data distributions is available, one can design parametric tests, such as Hotelling's two-sample test (Hotelling, 1931), Student's t-test (Pfanzagl and Sheynin, 1996), etc. Non-parametric two-sample tests are more attractive when the exact parametric form of the data distributions is hard to specify. It is popular to design non-parametric tests using integral probability metrics, since the evaluation of the corresponding test statistics can be obtained based on samples without knowing the densities of data distributions. Some earlier works design tests using Kolmogorov-Smirnov distance (Pratt and Gibbons, 1981; Jr., 1951), total variation distance (Györfi and Van Der Meulen, 1991), and Wasserstein distance (del Barrio et al., 1999; Ramdas et al., 2017). However, it is not proper to use these tests for high-dimensional settings since the sample complexity for estimating those distance functions based on empirical samples suffers from the curse of dimensionality.

There is a strong need for developing non-parametric tests for high-dimensional data, especially for modern applications. A notable contribution is the two-sample test based on Maximum Mean Discrepancy (MMD) (Gretton et al., 2009, 2012; Cheng and Xie, 2021a). Although the power of MMD test with the median choice of kernel bandwidth decays quickly when the dimension of distributions increases (Reddi et al., 2015), this test with properly chosen bandwidth does not have the curse of dimensionality issue for low-dimensional manifold data as pointed out in Cheng and Xie (2021a). Unfortunately, the MMD test with optimized bandwidth still does not demonstrate good testing power for the small-sampled case as demonstrated numerically in this paper. In addition, recent works (Wang et al., 2021; Xie and Xie, 2021) leverage the idea of dimensionality reduction for dealing with high-dimensional settings, which use the projected Wasserstein distance as the test statistic, i.e., the test statistic works by finding the linear projector such that

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

the distance between projected distributions is maximized. However, a linear projector may not serve as an optimal design for maximizing the power of tests as demonstrated numerically in Section 5.

In this paper, we present a new non-parametric two-sample test statistic aiming for the high-dimensional setting based on a *kernel projected Wasserstein (KPW) distance*, with a nonlinear projector based on the reproducing kernel Hilbert space (RKHS) designed to optimize the test power via maximizing the probability distance between the distributions after projection. In addition, our contributions include the following:

- We develop a computationally efficient algorithm for evaluating the KPW using a representer theorem to reformulate the problem into a finite-dimensional optimization problem and a block coordinate descent optimization algorithm which is guaranteed to find an ϵ-stationary point with complexity $\mathcal{O}\left(\epsilon^{-3}\right)$.
- To quantify the false detection rate, which is essential in setting the detection threshold, we develop non-asymptotic bounds for empirical KPW distance based on the covering number argument.
- We present numerical experiments to validate our theoretical results as well as demonstrate the competitive performance of our proposed test using both synthetic and real data.

Related Work. It is helpful to understand the structure of high-dimension distributions by low-dimensional projections. Notable methodologies include the principal component analysis (PCA) (Jolliffe, 1986), kernel PCA (Schölkopf et al., 1998), factor analysis (Cudeck, 2000), etc. Several works leverage this idea to design tests for high-dimensional data. Mueller and Jaakkola (2015) and Xie and Xie (2021) first design tests by finding the worst-case linear projector that maximizes the distance between projected sample points in one dimension. Later Lin et al. (2020) and Wang et al. (2021) naturally extend this idea by developing a projector that maps sample points into d dimensional linear subspace with $d \geq 1$, called projected Wasserstein distance. Efficient optimization algorithms and statistical properties of this distance have been investigated in recent works (Huang et al., 2021; Lin et al., 2021). However, a linear projector cannot efficiently capture features from data with nonlinear patterns, limiting the performance of tests mentioned above for practical applications. It is therefore promising to use nonlinear dimensionality reduction for two-sample testing. Although nonlinear projectors can be obtained using neural networks (Genevay et al., 2018), the sample complexity of the corresponding test statistic will have slow convergence rates since the neural network function class usually has high complexity in terms of the covering number. Recently kernel method has been demonstrated to be beneficial for understanding data (Minh and Sindhwani, 2011; Brouard et al., 2011; HQuang et al., 2013; Kadri et al., 2013) because of sharp sample complexity rate, low computational cost, and flexible representation of features. This fact motivates us to use a nonlinear projector based on kernels to design tests. Compared with the linear projector, computing the corresponding statistic and analyzing its performance is more challenging since the function space cannot be parameterized by finite-dimensional coefficients. We leverage the kernel trick to finish these two parts.

The remaining of this paper is organized as follows. Section 2 introduces some preliminary knowledge on two-sample testing and related probability distances, Section 3 outlines a practical algorithm for computing KPW distance, Section 4 studies the uncertainty quantification of empirical KPW distance, Section 5 demonstrates some numerical experiments, and Section 6 presents some concluding remarks.

2 PROBLEM SETUP

Let $x^n := \{x_i\}_{i=1}^n$ and $y^m := \{y_i\}_{i=1}^m$ be i.i.d. samples generated from distributions μ and ν supported on $\mathbb{R}^D$, respectively. Our goal is to design a two-sample test which, given samples x^n and y^m, decides to accept the null hypothesis $H_0 : \ \mu = \nu$ or reject H_0 in favor of the alternative hypothesis $H_1 : \ \mu \neq \nu$. Denote by $T : \ (x^n, y^m) \to \{t_0, t_1\}$ the two-sample test, where t_0 means we reject H_1 and t_1 means we accept H_1 and reject H_0. Define the type-I risk as the probability of rejecting hypothesis H_0 when it is true, and the type-II risk as the probability of accepting H_0 when $\mu \neq \nu$:

$$\epsilon_{n,m}^{(\mathrm{I})} = \mathbb{P}_{x^n \sim \mu, y^m \sim \nu}\Big(T(x^n, y^m) = t_1\Big), \quad \text{under } H_0,$$
$$\epsilon_{n,m}^{(\mathrm{II})} = \mathbb{P}_{x^n \sim \mu, y^m \sim \nu}\Big(T(x^n, y^m) = t_0\Big), \quad \text{under } H_1.$$

Given parameters $\alpha, \beta \in (0, \frac{1}{2})$, we aim at building a two-sample test such that, when applied to n-observation samples x^n and m-observation samples y^m, it has the type-I risk at most α (i.e., at level α) and the type-II risk at most β (i.e., of power $1 - \beta$). Moreover, we want to ensure these specifications with sample sizes n, m as small as possible.

We propose a non-parametric test by considering the probability distance functions between two empirical distributions constructed from observed samples. Specifically, we design a test T such that the null hy-

pothesis H_0 is rejected when

$$\mathcal{D}(\hat{\mu}_n, \hat{\nu}_m) > \chi,$$

where $\mathcal{D}(\cdot,\cdot)$ is a divergence quantifying the differences of two distributions, χ is a data-dependent threshold, and $\hat{\mu}_n$ and $\hat{\nu}_m$ are empirical distributions from n samples in μ and m samples in ν, respectively. Several existing tests can be unified into this framework by taking $\mathcal{D}(\cdot,\cdot)$ as some special probability distances, including the MMD test, total variation distance test, etc. In this paper, we will design the divergence $\mathcal{D}$ based on the Wasserstein distance, and we specify the cost function $c(x,y) = \|x-y\|_2^2$.

Definition 1 (Wasserstein Distance)**.** *Given two distributions μ and ν, the Wasserstein distance is defined as*

$$W(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,\mathrm{d}\pi(x,y),$$

where $c(\cdot,\cdot)$ denotes the cost function quantifying the distance between two points, and $\Pi(\mu,\nu)$ denotes the joint distribution with marginal distributions μ and ν.

Although Wasserstein distance has wide applications in machine learning, the finite-sample convergence rate of Wasserstein distance between empirical distributions is slow in high-dimensional settings (Fournier and Guillin, 2015). Therefore, it is not suitable for high-dimensional two-sample tests. Instead, existing works use the projection idea to rescue this issue.

Definition 2 (Projected Wasserstein Distance)**.** *Given two distributions μ and ν, define the projected Wasserstein distance as*

$$\mathcal{PW}(\mu,\nu) = \max_{\mathcal{A}:\ \mathbb{R}^D\to\mathbb{R}^d, A^{\mathrm{T}}A=I_d} W\left(\mathcal{A}\#\mu, \mathcal{A}\#\nu\right),$$

where the operator $\#$ denotes the push-forward operator, i.e.,

$$\mathcal{A}(z) \sim \mathcal{A}\#\mu \quad \text{for } z\sim\mu,$$

and we denote $\mathcal{A}$ as a linear operator such that $\mathcal{A}(z) = A^{\mathrm{T}}z$ with $z\in\mathbb{R}^D$ and $A\in\mathbb{R}^{D\times d}$.

This idea is demonstrated to be useful for breaking the curse of dimensionality for the original Wasserstein distance (Lin et al., 2021; Wang et al., 2021). However, a linear projector is not an optimal choice for dimensionality reduction. Instead, we will consider a nonlinear projector to obtain a more powerful two-sample test, and we use functions in vector-valued reproducing kernel Hilbert space (RKHS) for projection.

Definition 3 (Vector-valued RKHS)**.** *A function $K:\ \mathbb{R}^D\times\mathbb{R}^D\to\mathbb{R}^{d\times d}$ is said to be a positive semi-definite kernel if*

$$\sum_{i=1}^N\sum_{j=1}^N \langle \bar{y}_i, K(\bar{x}_i,\bar{x}_j)\bar{y}_j\rangle \geq 0$$

for any finite set of points $\{\bar{x}_i\}_{i=1}^N$ in $\mathbb{R}^D$ and $\{\bar{y}_i\}_{i=1}^N$ in $\mathbb{R}^d$. Given such a kernel, there exists an unique $\mathbb{R}^d$-valued Hilbert space $\mathcal{H}_K$ with the reproducing kernel K. For fixed $x\in\mathbb{R}^D$ and $y\in\mathbb{R}^d$, define the kernel section K_x with the action y as the mapping $K_xy:\ \mathbb{R}^D\to\mathbb{R}^d$ such that

$$(K_xy)(x') = K(x',x)y, \qquad \forall x'\in\mathbb{R}^D.$$

In particular, the Hilbert space $\mathcal{H}_K$ satisfies the reproducing property:

$$\forall f\in\mathcal{H}_K, \quad \langle f, K_xy\rangle_{\mathcal{H}_K} = \langle f(x), y\rangle.$$

Definition 4 (Kernel Projected Wasserstein Distance)**.** *Consider a $\mathbb{R}^d$-valued RKHS $\mathcal{H}$ with the corresponding kernel function K. Given two distributions μ and ν, define the kernel projected Wasserstein (KPW) distance as*

$$\mathcal{KPW}(\mu,\nu) = \max_{f\in\mathcal{F}} W\left(f\#\mu, f\#\nu\right)$$

where the function class $\mathcal{F} = \{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\leq 1\}$.

Remark 1. *For $d=1$, when the kernel function $K(x,y) = \langle x,y\rangle$, the KPW distance reduces into the PW distance. However, these two distances are not the same for general d. Moreover, existing works (Minh and Sindhwani, 2011; Micchelli and Pontil, 2005; Caponnetto et al., 2008; Baldassarre et al., 2010) consider the design of the matrix-valued kernel function for $d>1$ as*

$$K(x,x') = k(x,x')\cdot P, \tag{1}$$

where $k(\cdot,\cdot)$ denotes a scalar-valued kernel function and $P\in\mathbb{R}^{d\times d}$ is a positive semi-definite matrix that encodes the relation between the output space. Such a design reduces the computational cost for applying vector-valued RKHS.

In this paper, we design the two-sample test as follows. We split the data points into training and testing datasets. We first use the training set to train a nonlinear projector that maps data points into $\mathbb{R}^d$-subspace, and then perform the permutation test on testing data points that are projected based on the trained projector. The detailed algorithm is presented in Algorithm 1. This test is guaranteed to exactly control the type-I error (Good, 2013) because we evaluate the p-value of the test via the permutation approach. To obtain reliable two-sample tests, we also require the KPW distance satisfies the discriminative property that $\mathcal{KPW}(\mu,\nu) = 0$ if and only if $\mu=\nu$. The following proposition reveals that this property holds by considering the vector-valued RKHS satisfying the universal property, the proof of which is provided in Appendix C. We also study how to compute the kernel projected distance and its related statistical properties in the following sections.

Algorithm 1 Permutation two-sample test using the KPW distance

Require: Level α, number of permutation times N_p, collected samples x^n and y^m.

1: Split data as $x^n = x^{\mathrm{Tr}} \cup x^{\mathrm{Te}}$ and $y^m = y^{\mathrm{Tr}} \cup y^{\mathrm{Te}}$.
2: Formulate empirical distributions $(\hat{\mu}^{\mathrm{Tr}}, \hat{\nu}^{\mathrm{Tr}})$ corresponding to $(x^{\mathrm{Tr}}, y^{\mathrm{Tr}})$.
3: Obtain f as the (approximate) optimal projector to $\mathcal{KPW}(\hat{\mu}^{\mathrm{Tr}}, \hat{\nu}^{\mathrm{Tr}})$.
4: Compute the statistic $T = W(f\#\hat{\mu}^{\mathrm{Te}}, f\#\hat{\nu}^{\mathrm{Te}})$.
5: **for** $t = 1, \ldots, N_p$ **do**
6: Shuffle $x^{\mathrm{Te}} \cup y^{\mathrm{Te}}$ to obtain $x^{\mathrm{Te}}_{(t)}$ and $y^{\mathrm{Te}}_{(t)}$.
7: Formulate empirical distributions $(\hat{\mu}^{\mathrm{Te}}_{(t)}, \hat{\nu}^{\mathrm{Te}}_{(t)})$ corresponding to $(x^{\mathrm{Te}}, y^{\mathrm{Te}})$.
8: Compute the statistic for permuted samples $T_t = W(f\#\hat{\mu}^{\mathrm{Te}}_{(t)}, f\#\hat{\nu}^{\mathrm{Te}}_{(t)})$.
9: **end for**
Return the p-value $\frac{1}{N_p}\sum_{t=1}^{N_p} 1\{T_t \geq T\}$.

Proposition 1 (Discriminative Property of KPW)**.** *Denote by $\mathcal{C}_b(\mathcal{X})$ the space of bounded and continuous $\mathbb{R}^d$-valued functions on $\mathcal{X}$. Assume that $\mathcal{H}$ is a universal vector-valued RKHS so that for any $\varepsilon > 0$ and $f \in \mathcal{C}_b(\mathcal{X})$, there exists $g \in \mathcal{H}$ so that*

$$\|f - g\|_\infty \triangleq \sup_{x \in \mathcal{X}} \|f(x) - g(x)\|_2 < \varepsilon.$$

Then the KPW distance $\mathcal{KPW}(\mu, \nu) = 0$ if and only if $\mu = \nu$.

3 COMPUTING KPW DISTANCE

By the definition of Wasserstein distance, computing $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m)$ is equivalent to the following max-min problem:

$$\max_{f \in \mathcal{H}:\ \|f\|_{\mathcal{H}}^2 \leq 1} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} \|f(x_i) - f(y_j)\|_2^2 \right\}, \quad (2)$$

where $\Gamma = \left\{ \pi \in \mathbb{R}_+^{n \times m} : \ \sum_j \pi_{i,j} = \frac{1}{n}, \sum_i \pi_{i,j} = \frac{1}{m} \right\}$.

The computation of KPW distance has numerous challenges. It is crucial to design a suitable kernel function to obtain low computational complexity and reliable testing power, which will be discussed in Section 5. Moreover, the function $f \in \mathcal{H}$ is a countable combination of basis functions, i.e., the problem (2) is an infinite-dimensional optimization. By developing the representer theorem in Theorem 1, we are able to convert this problem into a finite-dimensional problem. Finally, there is no theoretical guarantee for finding the global optimum since it is a non-convex non-smooth optimization problem. Moreover, Sion's minimax theorem is not applicable because the problem (2) is not a convex programming: the inner minimization of quadratic function makes the objective in (2) not concave in f in general. Based on this observation, we only focus on optimization algorithms for finding a local optimum point in polynomial time.

Theorem 1 (Representer Theorem for KPW Distance)**.** *There exists an optimal solution to* (2) *that admits the following expression:*

$$\hat{f} = \sum_{i=1}^{n} K_{x_i} a_{x,i} - \sum_{j=1}^{m} K_{y_j} a_{y,j},$$

where $K_x(\cdot)$ denotes the kernel section and $a_{x,i}, a_{y,j} \in \mathbb{R}^d$ for $i = 1, \ldots, n, j = 1, \ldots, m$ are coefficients to be determined.

The proof of Theorem 1 is provided in Appendix D, in which standard representer theorem in literature (Schölkopf et al., 2001, Theorem 1) is not applicable since the RKHS norm serves as a hard constraint instead of the regularization of the objective function. In order to express the optimal solution as the compact matrix form, define $a_x \in \mathbb{R}^{nd}$ as the concatenation of coefficients $a_{x,i}$ for $i = 1, \ldots, n$ and

$$K_z(x^n) = \begin{pmatrix} K(z, x_1) & \cdots & K(z, x_n) \end{pmatrix} \in \mathbb{R}^{d \times nd}.$$

We also define the vector a_y and matrix $K_z(y^m)$ likewise. Then we have

$$\hat{f}(z) = K_z(x^n) a_x - K_z(y^m) a_y, \quad \forall z \in \mathcal{X}.$$

Define the gram matrix $K(x^n, x^n)$ as the $n \times n$ block matrix with the (i,j)-th block being $K(x_i, x_j)$. The gram matrices $K(x^n, y^m), K(y^m, x^n)$ and $K(y^m, y^m)$ can be defined likewise. Denote by G the concatenation of gram matrices:

$$G = \begin{pmatrix} K(x^n, x^n) & -K(x^n, y^m) \\ -K(y^m, x^n) & K(y^m, y^m) \end{pmatrix},$$

and we assume that G is positive definite. Otherwise, we add the gram matrix with a small number times identity matrix to make it invertible. Substituting the expression of $\hat{f}(z), z \in \mathcal{X}$ into (2), we obtain a finite-dimensional optimization problem:

$$\max_{\omega} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} : \ \omega^{\mathrm{T}} G \omega \leq 1 \right\},$$

where $\omega = [a_x^{\mathrm{T}}, a_y^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{d(n+m)}$, $c_{i,j} = \|A_{i,j}\omega\|_2^2$, and

$$A_{i,j} = [K_{x_i}(x^n) - K_{y_j}(x^n), K_{y_j}(y^m) - K_{x_i}(y^m)].$$

Suppose that the inverse of G admits the Cholesky decomposition $G^{-1} = UU^{\mathrm{T}}$, then by the change of

variable technique $s = U^{-1}\omega$, we obtain the norm-constrained optimization problem:

$$\max_{s\in\mathbb{R}^{d(n+m)}} \left\{ \min_{\pi\in\Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} : \ s^{\mathrm{T}} s \le 1 \right\}, \tag{3}$$

and we can replace the constraint $s^{\mathrm{T}}s \le 1$ with $s^{\mathrm{T}}s = 1$ based on the fact that the norm function satisfies the linear property. In other words, the decision variable s belongs to the Euclidean ball $\mathbb{S}^{d(n+m)-1} = \{s \in \mathbb{R}^{d(n+m)} : s^{\mathrm{T}}s = 1\}$.

For the ease of optimization, we consider the entropic regularization of the problem (3):

$$\max_{s\in\mathbb{S}^{d(n+m)-1}} \left\{ \min_{\pi\in\Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} - \eta H(\pi) \right\}, \tag{4}$$

in which we denote the entropy function $H(\pi) = -\sum_{i,j} \pi_{i,j}(\log \pi_{i,j} - 1)$. By the duality theory of entropic optimal transport (Genevay, 2019) and the change-of-variable technique, (4) is equivalent to the following minimization problem:

$$\min_{s\in\mathbb{S}^{d(n+m)-1}, u\in\mathbb{R}^n, v\in\mathbb{R}^m} F(u,v,s), \tag{5}$$

where

$$\begin{aligned} c_{i,j} &= \|A_{i,j}Us\|_2^2, \\ \pi_{i,j}(u,v,s) &= \exp\left(-\frac{1}{\eta}c_{i,j} + u_i + v_j\right), \\ F(u,v,s) &= \sum_{i,j} \pi_{i,j}(u,v,s) - \frac{1}{n}\sum_{i=1}^{n} u_i - \frac{1}{m}\sum_{j=1}^{m} v_j. \end{aligned}$$

The details for this deviation is deferred in Appendix D. Based on this formulation, we consider a Riemannian block coordinate descent (BCD) method (Hildreth, 1957) for optimization, which updates a block of variables by minimizing the objective function with respect to that block while fixing values of other blocks:

$$u^{t+1} = \min_{u\in\mathbb{R}^n} F(u, v^t, s^t), \tag{6a}$$

$$v^{t+1} = \min_{v\in\mathbb{R}^m} F(u^{t+1}, v, s^t), \tag{6b}$$

$$\zeta^{t+1} = \sum_{i,j} \nabla_s \pi_{i,j}(u^{t+1}, v^{t+1}, s^t), \tag{6c}$$

$$\xi^{t+1} = \mathcal{P}_{s^t}\big(\zeta^{t+1}\big), \tag{6d}$$

$$s^{t+1} = \mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big), \tag{6e}$$

where the operator $\mathcal{P}_s(\zeta)$ denotes the orthogonal projection of the vector ζ onto the tangent space of the manifold $\mathbb{S}^{d(n+m)-1}$ at s:

$$\mathcal{P}_s\big(\zeta\big) = \zeta - \langle s, \zeta\rangle s, \quad s \in \mathbb{S}^{d(n+m)-1},$$

Algorithm 2 BCD Algorithm for Solving (5)

Require: Empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_m$.
1: Initialize v^0, s^0
2: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
3: Update u^{t+1} according to (6g)
4: Update v^{t+1} according to (6h)
5: Update the Euclidean and Riemannian gradient ζ^{t+1} and ξ^{t+1}, according to (6i) and (6d), respectively.
6: Update s^{t+1} according to (6e)
7: **end for**
Return $u^* = u^T, v^* = v^T, s^* = s^T$.

and the retraction on this manifold is defined as

$$\mathrm{Retr}_s\big(-\tau\xi\big) = \frac{s - \tau\xi}{\|s - \tau\xi\|}, \quad s \in \mathbb{S}^{d(n+m)-1}. \tag{6f}$$

Note that the update steps (6a) and (6b) have closed-form expressions:

$$u^{t+1} = u^t + \left\{ \log \frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)} \right\}_{i\in[n]}, \tag{6g}$$

$$v^{t+1} = v^t + \left\{ \log \frac{1/m}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)} \right\}_{j\in[m]}, \tag{6h}$$

and the Euclidean gradient ζ^{t+1} in (6c) can be computed using the chain rule:

$$\zeta^{t+1} = -\frac{1}{\eta} U^{\mathrm{T}} \left[\sum_{i,j} \pi_{i,j}(u^{t+1}, v^{t+1}, s^t) A_{i,j}^{\mathrm{T}} A_{i,j} \right] U s^t. \tag{6i}$$

The overall algorithm for solving the problem (5) is summarized in Algorithm 2. We provide details for efficient implementation of the proposed algorithms in Appendix F. We also give a brief introduction to Riemannian optimization in Appendix B. The following theorem gives a convergence analysis of our proposed algorithm.The proof of this result is provided in Appendix D, which follows similar procedure in Huang et al. (2021). The main difference lies in establishing the descent lemma for updating the variable s on sphere instead of Stiefel manifold. Specifically, the procedure for finding the upper bound on the cost function $c_{i,j}$, the Lipschitz constant for $\pi_{i,j}(u,v,s)$ in s, and the Lipschitz constants of the retraction operator (6f) will be different.

Theorem 2 (Convergence Analysis for BCD)**.** *We say that $(\hat{u}, \hat{v}, \hat{s})$ is a (ϵ_1, ϵ_2)-stationary point of* (5) *if*

$$\begin{aligned} \|Grad_s F(\hat{u}, \hat{v}, \hat{s})\| &\le \epsilon_1, \\ F(\hat{u}, \hat{v}, \hat{s}) - \min_{u,v} F(u, v, \hat{s}) &\le \epsilon_2, \end{aligned}$$

where $Grad_s F(u,v,s)$ denotes the derivative of F with respect to s on the sphere $\mathbb{S}^{d(n+m)-1}$. Let $\{u^t, v^t, s^t\}$ be the sequence generated by Algorithm 2, then Algorithm 2 returns an (ϵ_1, ϵ_2)-stationary point in

$$T = \mathcal{O}\left(\log(mn) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2 \epsilon_2}\right]\right),$$

iterations, where the notation $O(\cdot)$ hides constants related to the initial guess (v^0, s^0) and the term $\max_{i,j} \|A_{i,j} U\|$.

Remark 2 (Complexity of Algorithm 2)**.** *Denote $N = n \vee m$[1]. Note that the iteration* (6g) *and* (6h) *can be implemented in $O(N)$ iterations. Second, the retraction step in* (6e) *requires $O(dN)$ arithmetic operations. Third, the computation of the Euclidean vector in* (6c) *can be implemented in $O(d^3N^3)$ operations, and the projection step can be done in $O(dN)$ operations. Therefore, the number of arithmetic operations in each iteration is of $O(d^3N^3)$. In summary, Algorithm 2 returns an (ϵ_1, ϵ_2)-stationary point in*

$$\mathcal{O}\left(d^3N^3 \log(N) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2 \epsilon_2}\right]\right)$$

arithmetic operations. Note that this computational complexity is independent of the dimension D of samples since we only need to compute the gram matrix as an input. The storage cost is of $\mathcal{O}(d^2N^2)$, in which the most expensive step is to store the gram matrix G.

4 PERFORMANCE GUARANTEES

In this section, we build statistical properties of the empirical KPW distance, though in practice we may not succeed in finding a global optimum solution to the non-convex optimization problem (2). We assume the cost function for the Wasserstein distance has the form $c(x,y) = \|x-y\|_2^p$ with $p \in [1, \infty)$. Moreover, results throughout this section are based on the following assumption.

Assumption 1. *For any $x, x' \in \mathcal{X}$, the matrix-valued kernel $K(x,x')$ is symmetric and satisfies*

$$0 \preceq K(x,x') \preceq BI_d.$$

Definition 5 ((Projection) Poincare Inequality)**.** *1. A distribution μ is said to satisfy a Poincare inequality if there exists an $M > 0$ for $X \sim \mu$ so that $Var[f(X)] \le M\mathbb{E}[\|\nabla f(X)\|^2]$ for any f satisfying $\mathbb{E}[f(X)^2] < \infty$ and $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$.*
2. A distribution μ is said to satisfy a projection Poincare inequality if there exists an $M > 0$ for any $f \in \mathcal{F}$ and $X \sim f\#\mu$ so that $Var[f(X)] \le M\mathbb{E}[\|\nabla f(X)\|^2]$ for any f satisfying $\mathbb{E}[f(X)^2] < \infty$ and $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$.

[1]We denote $a \vee b$ for $\max\{a,b\}$ and $a \wedge b$ for $\min\{a,b\}$.

Remark 3. *The Poincare inequality characterizes the relation about the variance of a function and its derivative in the spirit of the Sobolev inequality. It is a standard technical assumption for investigating the empirical convergence of Wasserstein distance (Lin et al., 2021; Lei, 2020), and is satisfied for various exponential measures such as the Gaussian distribution. See Ledoux (1999) for more examples.*

Lemma 1. *Assume that the distribution μ satisfies a projection Poincare inequality. Then*

$$\mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \lesssim n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p} + n^{-1/(2\vee p)} \sqrt{\log(n)} + n^{-1/p} \log(n),$$

where $\zeta_{p,d} = 1\{d = 2p\}$, and $\lesssim$ refers to "less than" with a constant depending only on (p, B).

Lemma 2. *Assume that the distribution μ satisfies a Poincare inequality, and any $f \in \mathcal{F}$ is L-Lipschitz. Then with probability at least $1-\alpha$, it holds that*

$$\left|(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}]\right| \le \max\left\{\varrho \log(1/\alpha), \sqrt{\varrho \log(1/\alpha)}\right\} n^{-1/(2\vee p)} L^{1/p},$$

where $\varrho > 0$ is a constant that depends on M.

Proof of two lemmas above follows similar covering number arguments in Lin et al. (2021), the details of which are deferred in Appendix E. The main difference is that we incorporate the reproducing property of vector-valued RKHS to give a valid bound on the covering number of the RKHS ball $\mathcal{F}$. Based on these two lemmas and the triangular inequality for Wasserstein distance, we give a finite-sample guarantee for the convergence of the KPW distance in Theorem 3. Compared with the sample complexity of estimating Wasserstein distance, KPW distance does not suffer from the curse of dimensionality as the RKHS ball $\mathcal{F}$ has low complexity.

Theorem 3 (Finite-sample Guarantee)**.** *Suppose the target distributions $\mu = \nu$, which satisfies projection Poincare inequality and Poincare inequality. Moreover, any $f \in \mathcal{F}$ is L-Lipschitz. Take $N = n \wedge m$, then with probability at least $1-2\alpha$, it holds that*

$$(\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m))^{1/p} \lesssim N^{-\frac{1}{(2p)\vee d}} (\log N)^{\zeta_{p,d}/p} + N^{-1/(2\vee p)} \sqrt{\log(N)} + N^{-1/p} \log(N) + \max\left\{\varrho \log(1/\alpha), \sqrt{\varrho \log(1/\alpha)}\right\} N^{-1/(2\vee p)} L^{1/p}.$$

4.1 Performance Guarantees for $p \in [1,2)$

When showing concentration results for p-Wasserstein distance with $p \in [1,2)$, however, it is not necessary to rely on the Poincare inequality assumption. The main result for this case is summarized in Theorem 4 (see details in Appendix E.3).

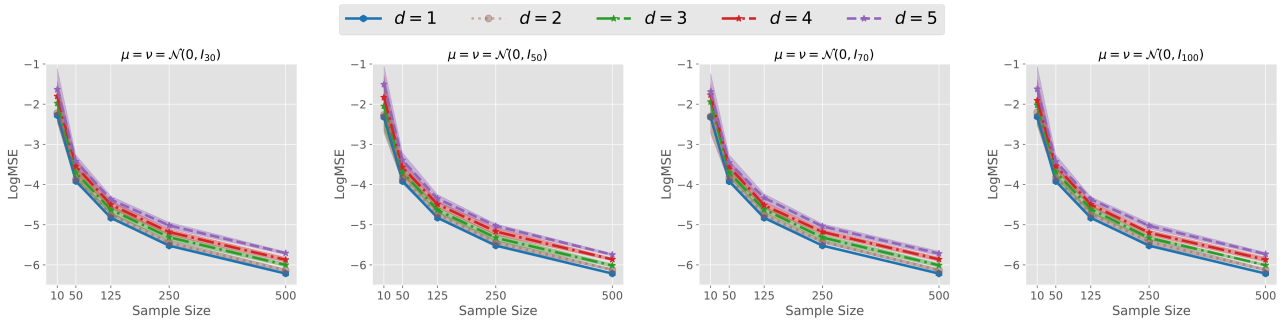

Figure 1: Average values of KPW distances between empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_n$ as the sample size n varies. Results are averaged for 10 independent trials and the shaded areas show the corresponding error bars.

Theorem 4 (Finite-sample Guarantee)**.** *Suppose the target distributions $\mu = \nu$. Then with probability at least $1 - 2\alpha$, it holds that*

$$\begin{aligned}(\mathcal{KPW}(\hat{\mu}_n, \nu_m))^{1/p} &\lesssim N^{-\frac{1}{(2p)\vee d}} (\log N)^{\zeta_{p,d}/p} \\ &+ N^{1/2-1/p}\sqrt{\log(N)} + N^{-1/p} \\ &+ N^{1/2-1/p}\sqrt{\log\frac{2}{\alpha}}.\end{aligned}$$

where $N = n \wedge m$ and $\lesssim$ refers to "less than" with a constant depending only on (p, B).

4.2 Sample Complexity

We also numerically examine the sample complexity of the empirical KPW distance $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_n)$ with $\mu = \nu = \mathcal{N}(0, I_D)$, where $n \in \{10, 50, 125, 250, 500\}$ and $D \in \{30, 50, 70, 100\}$. Figure 1 reports the average distances and the shaded areas show the corresponding error bars over 10 independent trials. We defer the detailed experiment setup and the plots of the computation time in Appendix G.1. From the plot we can see that the empirical KPW distances decay to zero quickly when the sample size n increases. Moreover, the distances with smaller values of d have faster decaying rates. Finally, the convergence behavior of the empirical KPW distances is nearly independent of the choice of D, which alleviates the issue of the curse of dimensionality for the original Wasserstein distance. These facts confirm the finite-sample guarantee discussed in Theorem 3.

5 NUMERICAL EXPERIMENTS

Throughout this section, we compare the performance of tests with the following procedures. (i) PW: the projected Wasserstein test where the projector is a linear mapping (Wang et al., 2021); (ii) MMD-O: the MMD test with a Gaussian kernel whose bandwidth is optimized (Liu et al., 2020); (iii) MMD-NTK: the test that combines both neural networks and MMD (Cheng and Xie, 2021b); and (iv) ME: the mean embedding test with optimized hyper-parameters (Jitkrittum et al., 2016). Implementation details on those baseline methods are omitted in Appendix G.2. When dealing with synthetic datasets, we generate a single sample set as the training set to learn parameters for each method. Then we evaluate the power of tests on 100 new sample sets generated from the same distribution. When dealing with real datasets, we randomly take part of samples as the training set, and evaluate the power on 100 randomly chosen subsets from the remaining samples. The number of permutations in Algorithm 1 is set to be $N_p = 100$. We control the type-I error for all tests at $\alpha = 0.05$.

When using the KPW distance, we follow (1) to design kernels to decrease the computational complexity. More specifically, we choose the scalar-valued kernel $k(\cdot, \cdot)$ to be a standard Gaussian kernel with the bandwidth σ^2, and

$$P = (1 - \rho)\mathbf{1}\mathbf{1}^{\mathrm{T}} + \rho I_d, \quad \text{with } \rho \in [0, 1].$$

We use the cross-validation approach to select the hyper-parameters ρ and σ^2, the details of which are deferred in Appendix G.3. The dimension d is prespecified and fixed into 3 in all experiments. We also present a study on the impact of hyper-parameters such as the projected dimension d and regularization parameter η in Appendix H.

5.1 Tests for Synthetic Datasets

We first investigate the performance when μ and ν are Gaussian distributions with diagonal covariance matrices. Specifically, we take $\mu = \mathcal{N}(0, I_D)$ and $\nu = \mathcal{N}(0, \Sigma)$ is the covariance shifted Gaussian, where the matrix $\Sigma = \mathrm{diag}(4, 4, 4, 1, \ldots, 1)$. In other words, we only scale the first three entries of the covariance matrix to make the high-dimensional testing problem challenging to handle. Fig. 2 reports the type-I and type-II errors for

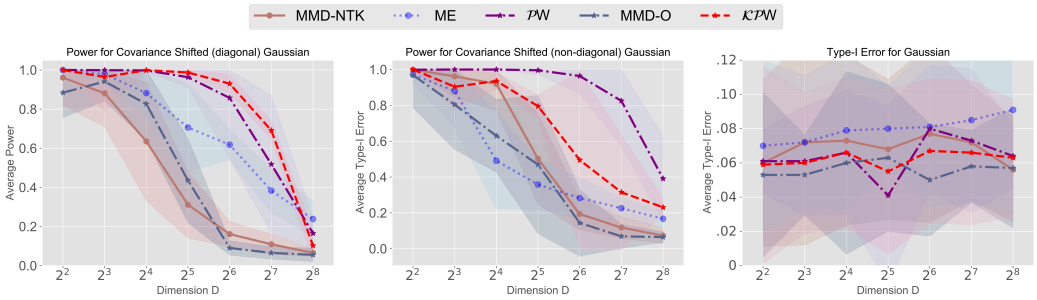

Figure 2: Testing results on Gaussian distributions across different choices of dimension D. Left: power for Gaussian distributions, where the shifted covariance matrix is still diagonal; Middle: power for Gaussian distributions, where the shifted covariance matrix is non-diagonal; Right: Type-I error.

Table 1: Average test power and standard error about detecting distribution abundance change in *MNIST* dataset across different choices of sample size.

N	MMD-NTK	MMD-O	ME	PW	KPW
200	0.639±0.029	**0.696±0.006**	0.298±0.031	0.302±0.033	0.663±0.015
250	0.763±0.010	0.781±0.002	0.472±0.017	0.369±0.030	**0.785±0.014**
300	0.813±0.016	0.869±0.002	0.630±0.025	0.524±0.023	**0.928±0.001**
400	0.881±0.013	0.956±0.003	0.779±0.020	0.591±0.044	**0.978±0.000**
500	0.950±0.002	0.988±0.000	0.927±0.006	0.782±0.040	**1.000±0.000**
Avg.	0.809	0.858	0.621	0.513	**0.870**

various tests across different choices of dimension D. We observe that both PW and KPW tests perform the best, while the power for other benchmark methods degrades quickly when the dimension D increases.

Next, we examine the case where ν has a non-diagonal covariance matrix. We take $\mu = \mathcal{N}(0, I_D)$ and $\nu = \mathcal{N}(0, V\Sigma V^{\mathrm{T}})$, where V is an orthogonal matrix with $V_{i,j} = \sqrt{2/(D+1)}\sin(ij\pi/(D+1))$ and $\Sigma = \mathrm{diag}(5, 5, 5, 1, \ldots, 1)$. Testing results for various choices of dimension D is reported in the middle of Fig. 2. In this case, the PW test performs slightly better than the KPW test. One possible explanation is that linear mapping seems to be the optimal choice for two-sample testing with covariance shifted Gaussian distributions. It is promising to design other types of matrix-valued kernel functions to improve performances of the KPW test.

Finally, we study the case where sample points are generated from high-dimensional Gaussian mixture distributions. We take $\mu = \frac{1}{2}\mathcal{N}(0, I_D) + \frac{1}{2}\mathcal{N}(\Delta_2, I_D)$ with $\Delta_2 = (1, 1, \ldots, 1)$ and $\nu = \frac{1}{2}\mathcal{N}(0, \Sigma_1) + \frac{1}{2}\mathcal{N}(\Delta_3, \Sigma_2)$ with $\Delta_3 = (1 + 0.8/\sqrt{D}, \ldots, 1 + 0.8/\sqrt{D})$. Covariance matrix Σ_1 is defined with $\Sigma_1[1,1] = \Sigma_1[2,2] = 4, \Sigma_1[1,2] = \Sigma_1[2,1] = -0.9, \Sigma_1[i,i] = 1, 3 \le i \le D$, and $\Sigma_1[i,j] = 0$ for indexes elsewhere. Covariance matrix Σ_2 is defined with $\Sigma_2[1,2] = \Sigma_2[2,1] = 0.9$, $\Sigma_2[i,i] = 1, 1 \le i \le D$, and $\Sigma_2[i,j] = 0$ for indexes elsewhere. Testing results (type-I and type-II errors) across different choices of dimension D for fixed sample size $n = m = 200$ is presented in the left two plots in Fig. 3. We also report results for increasing sample sizes $n = m$ by fixing the dimension $D = 140$ in the right two plots in Fig. 3. From the plot, we can see that all approaches have expected type-I error rates. Moreover, the tests based on PW and KPW distances outperform other benchmark methods, which indicates that the idea of dimension reduction is helpful for high-dimensional testing. The KPW test generally has the highest power in this case, since the nonlinear projector in the unit ball of RKHS is flexible enough to capture the differences between distributions. Other experiment details of this subsection is omitted in Appendix G.4.

5.2 Tests for MNIST handwritten digits

We now perform two-sample tests on the MNIST dataset (LeCun and Cortes, 2010). Let p be the distribution uniformly generated from the dataset, and $q = 0.85p + 0.15p_{\mathrm{cohort}}$, where p_{cohort} is the distribution from a class with digit 1. Both training and testing sample sizes are set to be $N \in \{200, 250, \ldots, 500\}$. Before performing two-sample tests, we pre-process this dataset by taking the sigmoid transformation of each image such that all scaled pixels are within the interval $[0, 1]$. Table 1 presents the testing power of various tests across different choices of N, from which we can see

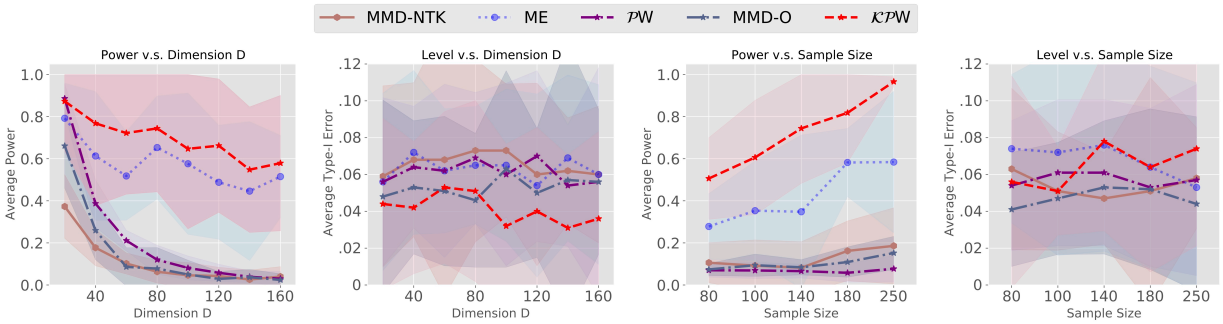

Figure 3: Testing results on Gaussian-mixture distributions. Left two: type-I and type-II errors across different choices of dimension D with fixed sample size $n = m = 200$; Right two: type-I and type-II errors across different choices of sample size $n = m$ with fixed dimension $D = 140$.

Table 2: Delay time for detecting the transition in *MSRC-12* that corresponds to four users.

User	MMD-NTK	MMD-O	ME	PW	KPW
1	36	73	82	47	**33**
2	8	7	97	9	**1**
3	15	13	27	**2**	20
4	22	83	69	16	**12**
Mean	20.25	44.0	68.8	18.50	**16.5**
Std	**12.0**	39.5	30.1	19.8	13.5

that the KPW test is competitive compared with other methods. We observe that performances of MMD-O in MNIST dataset are significantly better than that in synthetic datasets provided in Section 5.1. One possible explanation is that isotropic kernel functions will limit the power of MMD tests in some numerical examples (Liu et al., 2020, Section 3). Average type-I error for various tests is presented in Table 3 in Appendix G.5, from which we can see all tests have the type-I error close to $\alpha = 0.05$.

5.3 Human activity detection

Finally, we apply the KPW test to perform online change-point detection for human activity transition. We use a real-world dataset called the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture dataset (Fothergill et al., 2012). After pre-processing, this dataset consists of actions from four people, each with 855 samples in $\mathbb{R}^{60}$, and with a change of action from *bending* to *throwing* at the time index 500. More experimental details are omitted in Appendix G.6. Fix the window size $W = 100$. We pre-train a nonlinear projector using the data (sample size as the window) before time index 300 and compute the null statistics for many times to obtain the true threshold such that the false alarm rate is controlled within $\alpha = 0.05$. Then we perform online change-point detection based on a sliding window that moves forward with time. We compute the detection statistic by comparing the distribution between the block of data before time 300 and the data from the sliding window. We reject the null hypothesis and claim a change is happened if the statistic is above the threshold. Table 2 reports the delay time for detecting the behavior transition, from which we observe that the KPW test detects the change in the shortest time.

6 CONCLUSION

We proposed the KPW distance for the task of two-sample testing, which operates by finding the nonlinear mapping in the data space to maximize the distance between projected distributions. Practical algorithms together with uncertainty quantification of empirical estimates are discussed to help with this task.

The extension of this work is as follows. First, it is promising to consider milder technical assumptions than the projected Poincare inequality when establishing performance guarantees. Second, a meaningful research question is to determine the optimal hyper-parameters for the KPW test, including the projected subspace dimension d and the matrix-valued kernel function K. Third, it is desirable to study how to systematically pick the regularization parameter η to balance the trade-off between computational efficiency and accuracy of the obtained solution.

Acknowledgements

This work is supported by NSF DMS-2134037, CCF-1650913, CMMI-2015787, DMS-1938106, and DMS-1830210. The authors would like to thank the Editor and the anonymous referees for the thoughtful comments and suggestions, which led to an improvement of the presentation.

References

Ahmed, M., Mahmood, A. N., and Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31.

Baldassarre, L., Rosasco, L., Barla, A., and Verri, A. (2010). Vector field learning via spectral filtering. In *Machine Learning and Knowledge Discovery in Databases*, pages 56–71, Berlin, Heidelberg. Springer Berlin Heidelberg.

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

Boumal, N., Absil, P.-A., and Cartis, C. (2018). Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brouard, C., d'Alché Buc, F., and Szafranski, M. (2011). Semi-supervised penalized output kernel regression for link prediction. In *28th International Conference on Machine Learning (ICML 2011)*, pages 593–600.

Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, 9(52):1615–1646.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3).

Cheng, X. and Xie, Y. (2021a). Kernel mmd two-sample tests for manifold data. *arXiv preprint arXiv:2105.03425*.

Cheng, X. and Xie, Y. (2021b). Neural tangent kernel maximum mean discrepancy. In *Advances in Neural Information Processing Systems*, volume 34.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2606–2615.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

Cudeck, R. (2000). Exploratory factor analysis. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 265–296. Elsevier.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodriguez-Rodriguez, J. M. (1999). Tests of goodness of fit based on the l_2-wasserstein distance. *Annals of Statistics*, 27(4):1230–1239.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1737–1746. Association for Computing Machinery.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738.

Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1574–1583.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617.

Gin, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, USA.

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22, pages 673–681.

Györfi, L. and Van Der Meulen, E. C. (1991). *A Consistent Goodness of Fit Test Based on the Total Variation Distance*, pages 631–645. Springer Netherlands, Dordrecht.

Hildreth, C. (1957). A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85.

Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3):360–378.

HQuang, M., Bazzani, L., and Murino, V. (2013). A unifying framework for vector-valued manifold regularization and multi-view learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 100–108.

Hu, J., Liu, X., Wen, Z., and Yuan, Y. (2019). A brief introduction to manifold optimization. *arXiv preprint arXiv:1906.05450*.

Huang, M., Ma, S., and Lai, L. (2021). A riemannian block coordinate descent method for computing the projection robust wasserstein distance. *arXiv preprint arXiv:2012.05199*.

Jiang, B., Ma, S., So, A. M.-C., and Zhang, S. (2017). Vector transport-free svrg with general retraction for riemannian optimization: Complexity analysis and practical implementation. *arXiv preprint arXiv:1705.09059*.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016). Interpretable distribution features with maximum testing power. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 181–189.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag.

Jr., F. J. M. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Kadri, H., Rabaoui, A., Preux, P., Duflos, E., and Rakotomamonjy, A. (2013). Functional regularized least squares classi cation with operator-valued kernels. *arXiv preprint arXiv:1301.2655*.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Ledoux, M. (1999). Concentration of measure and logarithmic sobolev inequalities. *Séminaire de probabilités de Strasbourg*, 33:120–216.

Lei, J. (2020). Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1).

Lin, T., Fan, C., Ho, N., Cuturi, M., and Jordan, M. (2020). Projection robust wasserstein distance and riemannian optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 9383–9397.

Lin, T., Zheng, Z., Chen, E., Cuturi, M., and Jordan, M. (2021). On projection robust optimal transport: Sample complexity and model misspecification. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 262–270.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6316–6326.

Lloyd, J. R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837.

Lopez-Paz, D. and Oquab, M. (2018). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*.

McDiarmid, C. (1989). *On the method of bounded differences*, pages 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press.

Micchelli, C. A. and Pontil, M. A. (2005). On learning vector-valued functions. *Neural Computation*, 17(1):177–204.

Minh, H. Q. and Sindhwani, V. (2011). Vector-valued manifold regularization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 57–64.

Mueller, J. and Jaakkola, T. (2015). Principal differences analysis: Interpretable characterization of differences between distributions. In *Advances in Neural Information Processing Systems*, volume 28.

Pfanzagl, J. and Sheynin, O. (1996). Studies in the history of probability and statistics xliv a forerunner of the t-distribution. *Biometrika*, 83(4):891–898.

Poor, H. and Hadjiliadis, O. (2008). *Quickest detection*. Cambridge University Press.

Pratt, J. W. and Gibbons, J. D. (1981). *Kolmogorov-Smirnov Two-Sample Tests*, pages 318–344. Springer New York, New York, NY.

Ramdas, A., Garcia, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2).

Reddi, S. J., Ramdas, A., PAczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, page 3571–3577.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series. Princeton University Press.

Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014). Anomaly detection in online social networks. *Social networks*, 39:62–70.

Schober, P. and Vetter, T. (2019). Two-sample unpaired t tests in medical research. *Anesthesia and analgesia*, 129:911.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, pages 416–426.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, J., Gao, R., and Xie, Y. (2021). Two-sample test using projected wasserstein distance. In *Proceedings of IEEE International Symposium on Information Theory*.

Wen, Z. and Yin, W. (2012). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434.

Xie, L. and Xie, Y. (2021). Sequential change detection by optimal weighted ℓ_2 divergence. *IEEE Journal on Selected Areas in Information Theory*, pages 1–1.

Xie, L., Zou, S., Xie, Y., and Veeravalli, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2).

Young, G. A., Severini, T. A., Young, G. A., Smith, R., Smith, R. L., et al. (2005). *Essentials of statistical inference*, volume 16. Cambridge University Press.

Supplementary Material: Two-Sample Test with Kernel Projected Wasserstein Distance

A PRELIMINARY TECHNICAL RESULTS

Theorem 5 (Pinsker's Inequality (Cover and Thomas, 2006)). *Consider two discrete probability distributions $p = \{p_i\}_{i=1}^n$ and $q = \{q_i\}_{i=1}^n$, then it holds that*

$$\sum_{i=1}^n p_i \log \frac{p_i}{q_i} \geq \frac{1}{2}\|p - q\|_1^2.$$

Proposition 2 (Lipschitz Properties of Retraction Operator (Boumal et al., 2018)). *There exists constants L_1, L_2 such that the following inequalities hold:*

$$\|Retr_s(\zeta) - s\| \leq L_1\|\zeta\|$$
$$\|Retr_s(\zeta) - (s+\zeta)\| \leq L_2\|\zeta\|^2.$$

Inspired from Appendix A.3 in Jiang et al. (2017), we are able to compute the constants in Proposition 2 explicitly: $L_1 = 1$ and $L_2 = \frac{1}{2}$. The proof is provided below.

Proof. By definition, we have that

$$\begin{aligned}
\|\mathrm{Retr}_s(\zeta) - s\|_2^2 &= \left\| \frac{s+\zeta}{\|s+\zeta\|} - s \right\|_2^2 \\
&= 2\left(1 - \frac{1}{\|s+\zeta\|_2}\right) \\
&= 2\left(1 - (1 + \sum_i \zeta_i^2)^{-1/2}\right) \\
&\leq \sum_i \zeta_i^2 = \|\zeta\|_2^2.
\end{aligned}$$

where the second and the third equality is by using the relation $s^{\mathrm{T}}\zeta = 0$, and the inequality is based on the relation $2(1-(1+z)^{-1/2}) \leq z$ with $z = \sum_i \zeta_i^2$. Then it holds that $\|\mathrm{Retr}_s(\zeta) - (s+\zeta)\|_2 \leq \|\zeta\|$.

Secondly, we can see that

$$\begin{aligned}
\|\mathrm{Retr}_s(\zeta) - (s+\zeta)\|_2^2 &= \left\| \frac{s+\zeta}{\|s+\zeta\|} - (s+\zeta) \right\|_2^2 \\
&= \left(1 - \|s+\zeta\|_2\right)^2 \\
&= \left(1 - \sqrt{1 + \sum_i \zeta_i^2}\right)^2 \\
&\leq \frac{1}{4}\|\zeta\|_2^4,
\end{aligned}$$

where the inequality is based on the relation that $(1-(1+z)^{1/2})^2 \leq z^2/4$ with $z = \sum_i \zeta_i^2$. Consequently it holds that $\|\mathrm{Retr}_s(\zeta) - (s+\zeta)\|_2 \leq \frac{1}{2}\|\zeta\|^2$. □

Theorem 6 (McDiarmid's Inequality (McDiarmid, 1989)). *Let $X_1, \ldots, X_n$ be independent random variables, where X_i has the support $\mathcal{X}_i$. Let $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_n)$ bounded difference property, i.e., for $i \in \{1, \ldots, n\}$ and for any $(x_1, \ldots, x_n), (x_1', \ldots, x_n')$ that differs only in the i-th corodinate, we have*

$$|f(x_1, \ldots, x_n) - f(x_1', \ldots, x_n')| \leq c_i.$$

Then for any $t > 0$, we have

$$Pr\left\{ |f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t \right\} \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right).$$

Lemma 3 (Equivalent Definition for Sub-Gaussian variables (Lemma 2.3.2 in (Gin and Nickl, 2015))). *Assume that $\mathbb{E}[\zeta] = 0$ and*

$$\mathbb{P}\{|\zeta| \geq t\} \leq 2C \exp\left(-\frac{t^2}{2\sigma^2} \right), \quad t > 0,$$

for some $C \geq 1$ and $\sigma > 0$. Then the random variable ζ is sub-Gaussian with constant $\tilde{\sigma}^2 = 12(2C + 1)\sigma^2$.

Theorem 7 (Poincare's Inequality). *Denote by μ^n the product of μ on $\otimes_{i=1}^n \mathbb{R}^d$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies the Poincare's inequality, i.e., there exists $M > 0$ for $X \sim \mu$ so that $Var[f(X)] \leq M\mathbb{E}[\|\nabla f(X)\|^2]$ for any f satisfying $\mathbb{E}[f(X)^2] < \infty$ and $\mathbb{E}[\|\nabla f(X)\|_2^2] < \infty$. Consider a function f on $\otimes_{i=1}^n \mathbb{R}^d$ satisfying $\mathbb{E}|f(X)| < \infty$ and $\sum_{i=1}^n \|\nabla_i f(X)\|^2 \leq \alpha^2$, and $\max_{1 \leq i \leq n} \|\nabla_i f(X)\| \leq \beta$ almost surely. Then the following inequality holds for $X \sim \mu^n$:*

$$Pr\left\{ f(X) - \mathbb{E}[f(X)] > t \right\} \leq \exp\left(-\frac{1}{K} \min(t/\beta, t^2/\alpha^2) \right).$$

B INTRODUCTION TO MANIFOLD OPTIMIZATION

A brief introduction to manifold optimization can be found in Hu et al. (2019). In this section we list some related operators for solving manifold optimization problems. Traditional manifold optimization concerns with solving the following problem:

$$\min_{x\in\mathcal{M}} \ f(x), \tag{7}$$

where $\mathcal{M}$ is a Riemannian manifold and f is a real-valued function on $\mathcal{M}$. A tangent vector ζ_x to $\mathcal{M}$ at a point x is defined as a mapping so that there exists a curve γ on $\mathcal{M}$ satisfying

$$\gamma(0) = x, \quad \zeta_x[u] = \frac{\mathrm{d}(u(\gamma(t)))}{\mathrm{d}t}\mid_{t=0}, \ \forall u \in \mathfrak{E}(\mathcal{M}),$$

where $\mathfrak{E}(\mathcal{M})$ stands for the collection of real-valued functions defined in a neighborhood of x. Denote by $T_x\mathcal{M}$ as the collection of all tangent vectors to $\mathcal{M}$ at a point x, which is called the tangent space to $\mathcal{M}$ at x. Define $\mathcal{P}_x(z)$ as the projection of z into the tangent space at x. Based on definitions listed above, we can define necessary operators for manifold optimization. The Riemannian gradient of f at x is denoted as $\mathrm{Grad}f(x)$, which can be obtained by projecting the gradient of f at x in the Euclidean space into the tangent space to $\mathcal{M}$ at x:

$$\mathrm{Grad}f(x) = \mathcal{P}_x(\nabla f(x)).$$

Typical Riemannian manifolds include the Sphere and Stiefel manifold defined as follows:

$$\begin{aligned}\mathrm{Sphere}(n-1) &:= \{x \in \mathbb{R}^n : \ \|x\|_2 = 1\},\\ \mathrm{St}(n,p) &:= \{X \in \mathbb{R}^{n\times p} : \ X^{\mathrm{T}}X = I_p\}.\end{aligned}$$

We can express the tangent space together with the projection operator for these two types of manifolds in analytical form:

$$\begin{aligned}T_x\mathrm{Sphere}(n-1) &= \{z : \ z^{\mathrm{T}}x = 0\}, \quad \mathcal{P}_x(z) = (I - xx^{\mathrm{T}})z\\ T_x\mathrm{St}(n,p) &= \{Z : \ Z^{\mathrm{T}}X + X^{\mathrm{T}}Z = 0\}, \quad \mathcal{P}_X(Z) = Z - X\frac{X^{\mathrm{T}}Z + Z^{\mathrm{T}}X}{2}.\end{aligned}$$

When using first-order methods to solve a manifold optimization problem, one also needs to define the retraction operator associated with $\mathcal{M}$, which is denoted as Retr. It is a smooth mapping from the tangent budle $\cup_{x\in\mathcal{M}}T_x\mathcal{M}$ to $\mathcal{M}$ satisfying that for any $x \in \mathcal{M}$,

- $\mathrm{Retr}_x(0_x) = x$, where 0_x denotes the zero element in $T_x\mathcal{M}$;
- $\lim_{\zeta\in T_x\mathcal{M},\zeta\to 0} \frac{\|\mathrm{Retr}_x(\zeta)-(x+\zeta)\|}{\|\zeta\|} = 0$.

When $\mathcal{M}$ is a sphere, we choose the following retraction operator which can be implemented efficiently:

$$\mathrm{Retr}_x\big(\zeta\big) = \frac{x+\zeta}{\|x+\zeta\|}, \quad x \in \mathrm{Sphere}(n-1).$$

See Edelman et al. (1998) and Wen and Yin (2012) for discussions of retraction operators on the Stiefel manifold. The general iteration update of first-order methods for manifold optimization problem can be expressed as

$$x^{t+1} = \mathrm{Retr}_{x^t}(-\tau^t\zeta^t),$$

where τ^t is a well-defined step size and ζ^t is the Riemannian gradient at x^t. The computation of the projected Wasserstein distance relates to the optimization on a Stiefel manifold, while the computation of the KPW distance relates to the optimization on a sphere. A recent paper (Boumal et al., 2018) investigated the Riemannian gradient methods that are guaranteed to converge into stationary points globally, the key proof technique of which relies on Proposition 2. We follow the similar proof idea to establish the convergence analysis for computing the KPW distance.

C TECHNICAL PROOFS IN SECTION 2

Proof of Remark 1. When taking the kernel function $K(x, y) = \langle x, y\rangle$, the space

$$\mathcal{F} = \{a : \ a^{\mathrm{T}}a \leq 1\}.$$

Note that the cost function $c(x, y) = \|x - y\|_2^2$ satisfies $c(mx, my) = m^2 c(x, y)$ for any $m \in \mathbb{R}$. Hence we can argue that the maximizer of the KPW distance is obtained when $a^{\mathrm{T}}a = 1$, i.e.,

$$\mathcal{KPW}(\mu, \nu) = \max_{\substack{f:\ \mathbb{R}^D \to \mathbb{R}, \\ f(z) = a^{\mathrm{T}}z, a^{\mathrm{T}}a = 1}} W\left(f\#\mu, f\#\nu\right).$$

This indicates that the KPW distance reduces into the PW distance. □

Proof of Proposition 1. It is easy to see that $\mu = \nu$ implies $\mathcal{KPW}(\mu, \nu) = 0$. Now we show the converse. For fixed $x \in \mathcal{X}, y \in \mathbb{R}^d$ and a distribution μ, define the operator K_μ with the action y as a mapping $K_\mu y : \ \mathcal{X} \to \mathbb{R}^d$ so that

$$K_\mu y(x') = \int (K_x y)(x') \, \mathrm{d}\mu(x) = \int K(x', x) y \, \mathrm{d}\mu(x).$$

When $\mathcal{KPW}(\mu, \nu) = 0$, we can see that

$$f\#\mu = f\#\nu, \quad \forall f \in \mathcal{F},$$

which implies

$$\begin{aligned}
0 = & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \leq 1} \left\| \mathbb{E}_{f\#\mu}[x] - \mathbb{E}_{f\#\nu}[y] \right\|_2 \\
= & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \leq 1} \sup_{a:\ \|a\|_2 \leq 1} \left(\mathbb{E}_\mu[\langle f(x), a\rangle] - \mathbb{E}_\nu[\langle f(y), a\rangle] \right) \\
= & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \leq 1} \sup_{a:\ \|a\|_2 \leq 1} \left(\mathbb{E}_\mu[\langle f, K_x a\rangle_{\mathcal{H}}] - \mathbb{E}_\nu[\langle f, K_y a\rangle_{\mathcal{H}}] \right) \\
= & \sup_{f:\ \|f\|_{\mathcal{H}}^2 \leq 1} \sup_{a:\ \|a\|_2 \leq 1} \langle f, (K_\mu - K_\nu) a\rangle \\
= & \sup_{a:\ \|a\|_2 \leq 1} \|(K_\mu - K_\nu) a\|_{\mathcal{H}}.
\end{aligned}$$

Equivalently, $\|(K_\mu - K_\nu)a\|_{\mathcal{H}} = 0$ for any a so that $\|a\|_2 \leq 1$. Since $\mathcal{H}$ is a Hilbert space, we imply that $(K_\mu - K_\nu)a$ is a zero function for any a satisfying $\|a\|_2 \leq 1$. For any function $f \in \mathcal{C}(X)$, we make the expansion

$$\begin{aligned}
& \left\| \mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)] \right\|_2 \\
\leq & \left\| \mathbb{E}_\mu[f(x)] - \mathbb{E}_\mu[g(x)] \right\|_2 + \left\| \mathbb{E}_\mu[g(x)] - \mathbb{E}_\nu[g(y)] \right\|_2 + \left\| \mathbb{E}_\nu[g(y)] - \mathbb{E}_\nu[f(y)] \right\|_2.
\end{aligned}$$

The first term satisfies that

$$\left\| \mathbb{E}_\mu[f(x)] - \mathbb{E}_\mu[g(x)] \right\|_2 \leq \mathbb{E}_\mu[\|f(x) - g(x)\|_2] < \varepsilon,$$

and the third term can be upper bounded likewise. For the second term, we have that

$$\begin{aligned}
& \left\| \mathbb{E}_\mu[g(x)] - \mathbb{E}_\nu[g(y)] \right\|_2 \\
= & \sup_{a:\ \|a\|_2 \leq 1} \left(\mathbb{E}_\mu[\langle g(x), a\rangle] - \mathbb{E}_\nu[\langle g(y), a\rangle] \right) \\
= & \sup_{a:\ \|a\|_2 \leq 1} \left(\mathbb{E}_\mu[\langle g, K_x a\rangle] - \mathbb{E}_\nu[\langle g, K_y a\rangle] \right) \\
= & \sup_{a:\ \|a\|_2 \leq 1} \langle g, (K_\mu - K_\nu) a\rangle = 0,
\end{aligned}$$

where the last equality is because that $(K_\mu - K_\nu)a$ is a zero function for any a satisfying $\|a\|_2 \leq 1$. Hence, $\left\| \mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)] \right\|_2 < 2\varepsilon$ for any $\varepsilon > 0$ and $f \in \mathcal{C}_b(\mathcal{X})$. Then we conclude that the distribution $\mu = \nu$. □

D TECHNICAL PROOFS IN SECTION 3

D.1 Deviation of Duality Reformulation (5)

We first present the proof of the dual reformulation of the inner minimization problem in (4). By definition, the primal formulation can be expressed as:

$$\min_{\pi\geq 0} \left\{ \sum_{i,j} \pi_{i,j} c_{i,j} - \eta \sum_{i,j} \pi_{i,j}(\log \pi_{i,j} - 1) : \quad \sum_j \pi_{i,j} = \frac{1}{n}, \sum_i \pi_{i,j} = \frac{1}{m} \right\}. \tag{8}$$

The Lagrangian function becomes

$$L(\pi, u, v) = \sum_{i,j} \pi_{i,j} c_{i,j} - \eta \sum_{i,j} \pi_{i,j}(\log \pi_{i,j} - 1) + \sum_i u_i \left(\sum_j \pi_{i,j} - \frac{1}{n} \right) + \sum_j v_j \left(\sum_i \pi_{i,j} - \frac{1}{m} \right)$$

Then the dual problem becomes

$$\begin{aligned}
&\max_{u,v} \ \left\{ \min_{\pi\geq 0} \ L(\pi, u, v) \right\} \\
=&\max_{u,v} \ -\frac{1}{n}\sum_i u_i - \frac{1}{m}\sum_j v_j + \min_{\pi\geq 0} \ \sum_{i,j} \pi_{i,j}\big[c_{i,j} + u_i + v_j\big] - \eta \pi_{i,j}(\log \pi_{i,j} - 1) \\
=&\max_{u,v} \ -\frac{1}{n}\sum_i u_i - \frac{1}{m}\sum_j v_j - \sum_{i,j} \max_{\pi_{i,j}\geq 0} \ \big\{ -\pi_{i,j}\big[c_{i,j} + u_i + v_j\big] + \eta \pi_{i,j}(\log \pi_{i,j} - 1) \big\} \\
=&\max_{u,v} \ -\frac{1}{n}\sum_i u_i - \frac{1}{m}\sum_j v_j - \sum_{i,j} (\eta\phi)^*(u_i + v_j + c_{i,j}) \\
=&\max_{u,v} \ -\frac{1}{n}\sum_i u_i - \frac{1}{m}\sum_j v_j - \eta \sum_{i,j} \exp\left(-\frac{u_i + v_j + c_{i,j}}{\eta} \right)
\end{aligned}$$

where $\phi(w) = w\log w - w$ and ϕ^* denotes its conjugate (Rockafellar, 1970). Moreover, the dual optimal value equals the primal optimal value because the Slater's condition (Boyd and Vandenberghe, 2004) for finite-dimensional optimization is satisfied. Take $u_i' = -u_i/\eta$ and $v_j' = -v_j/\eta$, the dual problem becomes

$$\max_{u',v'} \ \frac{\eta}{n}\sum_i u_i' + \frac{\eta}{m}\sum_j v_j' - \eta \sum_{i,j} \exp\left(-\frac{c_{i,j}}{\eta} + u_i' + v_j' \right).$$

Therefore, the whole problem (4) becomes

$$\max_{u,v,s} \ \frac{\eta}{n}\sum_i u_i + \frac{\eta}{m}\sum_j v_j - \eta \sum_{i,j} \exp\left(-\frac{c_{i,j}}{\eta} + u_i + v_j \right).$$

Or equivalently, we write it as the minimization problem:

$$-\eta \times \left\{ \min_{u,v,s} \ -\frac{1}{n}\sum_i u_i - \frac{1}{m}\sum_j v_j + \eta \sum_{i,j} \exp\left(-\frac{c_{i,j}}{\eta} + u_i + v_j \right) \right\}.$$

Remark 4. *By adding the entropic regularization term $\eta H(\pi)$, we are able to derive an unconstrained optimization formulation on the sphere, thus reducing the computational cost for computing KPW distance. Besides, the induced optimal transport mapping between projected samples is usually stochastic instead of deterministic, which is robust to potential data outliers.*

D.2 Proof of Theorem 1

Assume that $\hat{f}$ is an optimal solution to the problem (2). Let S be the subspace

$$S = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} (K_{x_i} - K_{y_j}) a_{i,j} : \ a_{i,j} \in \mathbb{R}^d \right\}.$$

Denote by $S_\perp$ the orthogonal complement of S. Given a set $\mathcal{X}$, denote by $f_{\mathcal{X}}$ a function that lies in the set $\mathcal{X}$. Then by the projection theorem, there exists $\hat{f}_S$ and $\hat{f}_{S_\perp}$ such that $\hat{f} = \hat{f}_S + \hat{f}_{S_\perp}$ and $\|\hat{f}\|_{\mathcal{H}}^2 = \|\hat{f}_S\|_{\mathcal{H}}^2 + \|\hat{f}_{S_\perp}\|_{\mathcal{H}}^2$. It remains to show that $\hat{f}_S$ shares the same objective value with $\hat{f}$. For fixed i, j, we have that

$$\begin{aligned}
\|\hat{f}(x_i) - \hat{f}(y_j)\|_2 &= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}(x_i) - \hat{f}(y_j), a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}(x_i), a_{i,j} \rangle - \langle \hat{f}(y_j), a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}, K_{x_i} a_{i,j} \rangle - \langle \hat{f}, K_{y_j} a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}, (K_{x_i} - K_{y_j}) a_{i,j} \rangle \\
&= \max_{a_{i,j}:\ \|a_{i,j}\|_2 \le 1} \langle \hat{f}_S, (K_{x_i} - K_{y_j}) a_{i,j} \rangle = \|\hat{f}_S(x_i) - \hat{f}_S(y_j)\|_2,
\end{aligned}$$

where the second last equality is because $\hat{f}_{S_\perp}$ is orthogonal to the subspace S. It follows that $\|\hat{f}(x_i) - \hat{f}(y_j)\|_2^2 = \|\hat{f}_S(x_i) - \hat{f}_S(y_j)\|_2^2$. Therefore, there always exists an optimal solution that lies in the subspace S, which means that there exists an optimal solution to (2) that admits the following expression:

$$\hat{f} = \sum_{i=1}^{n} \sum_{j=1}^{m} (K_{x_i} - K_{y_j}) a_{i,j}.$$

Defining $a_{x,i} = \sum_{j=1}^{m} a_{i,j}$ and $a_{y,j} = \sum_{i=1}^{n} a_{i,j}$ completes the proof.

Remark 5. *From the proof we can also see that the representer theorem holds if replacing the square of the ℓ_2 norm in (2) with any p-th power of the ℓ_2 norm for $p \ge 2$. However, we find the development of optimization algorithms for the square of the ℓ_2 norm case is the simplest.*

D.3 Proof of Theorem 2

In the following we give a iteration complexity analysis about Algorithm 2, the proof of which largely follows the idea in Huang et al. (2021). In particular, we first establish the descent lemma for the update of each block of variables and then argue that the objective function is lower bounded. Based on these two facts, we finally build the iteration complexity result for Algorithm 2.

Lemma 4 (Lipschitzness of $\nabla_s F(u, v, s)$)**.** *Let $\{u^t, v^t, s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $s \in \mathbb{S}^{d(n+m)-1}$ and $\lambda \in [0, 1]$:*

$$\|\nabla_s F(u^{t+1}, v^{t+1}, \lambda s + (1-\lambda) s^t) - \nabla_s F(u^{t+1}, v^{t+1}, s^t)\| \le \varrho \lambda \|s^t - s\|,$$

where $\varrho = \frac{2\|AU\|_\infty^2}{\eta} + \frac{4\|AU\|_\infty^4}{\eta^2}$ and $\|AU\|_\infty = \max_{i,j} \|A_{i,j} U\|_2$.

Proof of Lemma 4. An intermediate result is that

$$\begin{aligned}
\sum_i \pi_{i,j}(u^{t+1}, v^{t+1}, s^t) &= \sum_i \exp\left(-\frac{1}{\eta} c_{i,j}[s^t] + u_i^{t+1} \right) \exp\left(v_j^{t+1}\right) \\
&= \sum_i \exp\left(-\frac{1}{\eta} c_{i,j}[s^t] + u_i^{t+1} \right) \exp\left(v_j^{t}\right) \frac{1/m}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)} \\
&= \frac{1}{m} \frac{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)}{\sum_i \pi_{i,j}(u^{t+1}, v^t, s^t)} = 1/m.
\end{aligned}$$

Then we can assert that $\sum_{i,j}\pi_{i,j}(u^{t+1},v^t,s^t)=1$. For fixed s^t, define $s^\lambda=\lambda s+(1-\lambda)s^t$. Then we have that

$$\begin{aligned}
&\|\nabla_s F(u^{t+1},v^{t+1},s^t)-\nabla_s F(u^{t+1},v^{t+1},s^\lambda)\|\\
=&\frac{2}{\eta}\left\|\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}Us^t-\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}Us^\lambda\right\|\\
\le&\frac{2}{\eta}\left\|\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U(s^t-s^\lambda)\right\|\\
&+\frac{2}{\eta}\left\|\sum_{i,j}U^{\mathrm{T}}\big[\pi_{i,j}(u^{t+1},v^{t+1},s^t)-\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)\big]A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\\
\le&\frac{2}{\eta}\left\|\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\|s^\lambda-s^t\|\\
&+\frac{2}{\eta}\left\|\sum_{i,j}\big[\pi_{i,j}(u^{t+1},v^{t+1},s^t)-\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)\big]U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|
\end{aligned}$$

where the first inequality is based on the constraint that $\|s^\lambda\|\le\lambda\|s\|+(1-\lambda)\|s^t\|=1$. To upper bound the first term, we find

$$\begin{aligned}
&\left\|\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\\
\le&\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)\|U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\|_2\le\max_{i,j}\|A_{i,j}U\|_2^2.
\end{aligned}$$

To bound the second term, we find that

$$\begin{aligned}
&\left\|\sum_{i,j}\big[\pi_{i,j}(u^{t+1},v^{t+1},s^t)-\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)\big]U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\\
\le&\max_{i,j}\|A_{i,j}U\|_2^2\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1,
\end{aligned}$$

where

$$\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1:=\sum_{i,j}\big|\pi_{i,j}(u^{t+1},v^{t+1},s^\lambda)-\pi_{i,j}(u^{t+1},v^{t+1},s^t)\big|.$$

Denote by $H(\pi,s;\eta)$ the objective function for (3). Based on the strong convexity property, we have that

$$\begin{aligned}
&\langle\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^\lambda),s^\lambda;\eta)-\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^\lambda;\eta),\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\rangle\\
\ge&\eta\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1^2
\end{aligned}$$

Moreover, by simple calculation we find

$$\begin{aligned}
\nabla_\pi H(\pi(u,v,s),s)&=[c_{i,j}+\eta\log(\pi_{i,j}(u,v,s))]_{i,j}\\
&=[\eta(u_i+v_j)]_{i,j},
\end{aligned}$$

where the second equality is by substituting the formulation of $\pi_{i,j}(u,v,s)$. Hence, we find that the gradient $\nabla_\pi H(\pi(u,v,s),s)$ only depends on u and v, which implies

$$\begin{aligned}
&\langle\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^t;\eta)-\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^t),s^\lambda;\eta),\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\rangle\\
\ge&\eta\|\pi(u^{t+1},v^{t+1},s^\lambda)-\pi(u^{t+1},v^{t+1},s^t)\|_1^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\eta\|\pi(u^{t+1},v^{t+1},s^{\lambda})-\pi(u^{t+1},v^{t+1},s^{t})\|_1\\
\leq&\|\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^{t}),s^{t};\eta)-\nabla_\pi H(\pi(u^{t+1},v^{t+1},s^{t}),s^{\lambda};\eta)\|_\infty\\
=&\max_{i,j}\left|\|A_{i,j}Us^{\lambda}\|_2^2-\|A_{i,j}Us^{t}\|_2^2\right|\\
\leq&2\max_{i,j}\|A_{i,j}U\|_2^2\|s^{\lambda}-s^{t}\|.
\end{aligned}$$

where the inequality is by applying the following relation:

$$\begin{aligned}
\|Ax_1\|_2^2-\|Ax_2\|_2^2&=(x_1-x_2)^{\mathrm{T}}(A^{\mathrm{T}}Ax_1)+x_2^{\mathrm{T}}A^{\mathrm{T}}A(x_1-x_2)\\
&\leq\|x_1-x_2\|\|A^{\mathrm{T}}Ax_1\|+\|x_2^{\mathrm{T}}A^{\mathrm{T}}A\|\|x_1-x_2\|\\
&\leq 2\|A\|^2\|x_1-x_2\|.
\end{aligned}$$

In summary, the second term can be upper bounded as

$$\begin{aligned}
&\left\|\sum_{i,j}\left[\pi_{i,j}(u^{t+1},v^{t+1},s^{t})-\pi_{i,j}(u^{t+1},v^{t+1},s^{\lambda})\right]U^{\mathrm{T}}A_{i,j}^{\mathrm{T}}A_{i,j}U\right\|\\
\leq&\frac{2\left(\max_{i,j}\|A_{i,j}U\|_2^2\right)^2}{\eta}\|s^{\lambda}-s^{t}\|.
\end{aligned}$$

Then applying the condition that $\|s^{\lambda}-s^{t}\|=\lambda\|s-s^{t}\|$ completes the proof. □

Lemma 5 (Decrease of F in s)**.** *Let $\{u^t,v^t,s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $k\geq 1$:*

$$F(u^{t+1},v^{t+1},s^{t+1})-F(u^{t+1},v^{t+1},s^{t})\leq-\frac{1}{8\|AU\|_\infty^2L_2/\eta+2\varrho L_1^2}\|\xi^{t+1}\|^2.$$

Proof of Lemma 5. Note that

$$\begin{aligned}
&\left|F(u^{t+1},v^{t+1},s^{t+1})-F(u^{t+1},v^{t+1},s^{t})-\langle\nabla_tF(u^{t+1},v^{t+1},s^{t}),s^{t+1}-s^{t}\rangle\right|\\
=&\left|\int_0^1\langle\nabla_sF(u^{t+1},v^{t+1},\lambda s^{t+1}+(1-\lambda)s^{t})-\nabla_sF(u^{t+1},v^{t+1},s^{t}),s^{t+1}-s^{t}\rangle\,\mathrm{d}\lambda\right|\\
\leq&\int_0^1\|\nabla_sF(u^{t+1},v^{t+1},\lambda s^{t+1}+(1-\lambda)s^{t})-\nabla_sF(u^{t+1},v^{t+1},s^{t})\|\|s^{t+1}-s^{t}\|\,\mathrm{d}\lambda\\
\leq&\int_0^1\varrho\lambda\|s^{t+1}-s^{t}\|^2\,\mathrm{d}\lambda\\
=&\frac{\varrho}{2}\|s^{t+1}-s^{t}\|^2=\frac{\varrho}{2}\left\|\mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big)-s^{t}\right\|^2\\
\leq&\frac{\varrho\tau^2L_1^2}{2}\|\xi^{t+1}\|^2.
\end{aligned}$$

where the second inequality is by applying Lemma 4, and the last inequality is by applying Proposition 2. Moreover, we have that

$$\begin{aligned}
&\langle\nabla_sF(u^{t+1},v^{t+1},s^{t}),s^{t+1}-s^{t}\rangle\\
=&\langle\nabla_sF(u^{t+1},v^{t+1},s^{t}),-\tau\xi^{t+1}\rangle+\langle\nabla_sF(u^{t+1},v^{t+1},s^{t}),\mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big)-(s^{t}-\tau\xi^{t+1})\rangle\\
\leq&-\tau\|\xi^{t+1}\|^2+\|\nabla_sF(u^{t+1},v^{t+1},s^{t})\|_2\|\mathrm{Retr}_{s^t}\big(-\tau\xi^{t+1}\big)-(s^{t}-\tau\xi^{t+1})\|\\
\leq&-\tau\|\xi^{t+1}\|^2+\|\zeta^{t+1}\|_2\cdot L_2\tau^2\|\xi^{t+1}\|^2\\
\leq&-\tau\|\xi^{t+1}\|^2+\frac{2\|AU\|_\infty^2L_2\tau^2}{\eta}\|\xi^{t+1}\|^2.
\end{aligned}$$

Combining those inequalities above implies that

$$F(u^{k+1},v^{k+1},t^{k+1})-F(u^{k+1},v^{k+1},t^{k})\le-\tau\left(1-\left[\frac{2\|AU\|_\infty^2L_2}{\eta}+\frac{\varrho}{2}L_1^2\right]\tau\right)\|\xi^{t+1}\|^2.$$

Taking $\tau=\frac{1}{4\|AU\|_\infty^2L_2/\eta+\varrho L_1^2}$ gives the desired result. □

Lemma 6 (Decrease of F in v)**.** *Let $\{u^t,v^t,s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $k\ge1$:*

$$F(u^{t+1},v^{t+1},s^t)-F(u^{t+1},v^t,s^t)\le-\frac{1}{2}\|1/m-\pi(u^{t+1},v^t,s^t)^{\mathrm{T}}1\|_1^2.$$

where

$$\|1/m-\pi(u^{t+1},v^t,s^t)\|_1=\sum_j\left|\frac{1}{m}-\sum_i\pi_{i,j}(u^{t+1},v^t,s^t)\right|.$$

Proof of Lemma 6. According to the expression of F, we have that

$$\begin{aligned}&F(u^{t+1},v^{t+1},s^t)-F(u^{t+1},v^t,s^t)\\=&\sum_{i,j}\pi_{i,j}(u^{t+1},v^{t+1},s^t)-\sum_{i,j}\pi_{i,j}(u^{t+1},v^t,s^t)+\frac{1}{m}\sum_{j=1}^m(v_j^t-v_j^{t+1})\\=&\frac{1}{m}\sum_{j=1}^m(v_j^t-v_j^{t+1})=-\frac{1}{m}\sum_{j=1}^m\log\frac{1/m}{\sum_i\pi_{i,j}(u^{t+1},v^t,s^t)},\end{aligned}$$

where the second equality is because that

$$\sum_i\pi_{i,j}(u^{t+1},v^{t+1},s^t)=\frac{1}{m},$$
$$\sum_j\pi_{i,j}(u^{t+1},v^t,s^t)=\frac{1}{n}.$$

Therefore, applying the Pinsker's inequality in Theorem 5 implies that

$$F(u^{t+1},v^{t+1},s^t)-F(u^{t+1},v^t,s^t)\le-\frac{1}{2}\left(\sum_j\left|\frac{1}{m}-\sum_i\pi_{i,j}(u^{t+1},v^t,s^t)\right|\right)^2.$$

□

Lemma 7 (Decrease of F in u)**.** *Let $\{u^t,v^t,s^t\}_t$ be the sequence generated from Algorithm 2. The following inequality holds for any $t\ge1$:*

$$F(u^{t+1},v^t,s^t)-F(u^t,v^t,s^t)\le-\frac{1}{2}\|1/n-\pi(u^t,v^t,s^t)1\|_2^2.$$

where

$$\|1/n-\pi(u^t,v^t,s^t)1\|_2^2=\sum_i\left|\frac{1}{n}-\sum_j\pi_{i,j}(u^k,v^k,t^k)\right|^2.$$

Proof of Lemma 7. For fixed $i\in[n]$, define

$$h_i=\sum_j\pi_{i,j}(u^{t+1},v^t,s^t)-\sum_j\pi_{i,j}(u^t,v^t,s^t)-\frac{1}{n}\log\frac{1/n}{\sum_j\pi_{i,j}(u^t,v^t,s^t)}$$

According to the expression of F,

$$F(u^{t+1}, v^t, s^t) - F(u^t, v^t, s^t) = \sum_i h_i,$$

and it suffices to provide an upper bound for $h_i, i \in [n]$. By substituting the expression of u^{t+1} into h_i, we have that

$$\begin{aligned} h_i &= \sum_j \pi_{i,j}(u^t, v^t, s^t) \left[\frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)} - 1 \right] - \frac{1}{n} \log \frac{1/n}{\sum_j \pi_{i,j}(u^t, v^t, s^t)} \\ &= \frac{1}{n} - \big(\pi(u^t, v^t, s^t)1\big)_i - \frac{1}{n} \log \frac{1/n}{\big(\pi(u^t, v^t, s^t)1\big)_i} \end{aligned}$$

Define the function

$$\ell(x) = \frac{1}{n} - x - \frac{1}{n} \log \frac{1/n}{x} + (x - 1/n)^2.$$

We can see that this function attains its maximum at $x = 1/n$, with $\ell(1/n) = 0$. It follows that

$$h_i \le -\left(\big(\pi(u^t, v^t, s^t)1\big)_i - \frac{1}{n} \right)^2.$$

The proof is completed. □

Lemma 8. *Let $\{u^t, v^t, s^t\}_t$ be the sequence generated from Algorithm 2, which is terminated when the following conditions hold:*

$$\|\xi^{t+1}\| \le \epsilon_1, \quad \|1/n - \pi(u^t, v^t, s^t)1\|_2 \le \frac{\epsilon_2}{4\|AU\|_\infty^2}, \quad \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1 \le \frac{\epsilon_2}{4\|AU\|_\infty^2}.$$

Then $\{u^T, v^T, s^T\}$ is an (ϵ_1, ϵ_2) stationary point of (5).

Proof of Lemma 8. The condition $\|\xi^{t+1}\| \le \epsilon_1$ directly implies that

$$\|\mathrm{Grad}_s F(u^T, v^T, s^T)\| \le \epsilon_1.$$

Suppose that

$$\pi(u^T, v^T, s^T)1 = r, \quad \pi(u^T, v^T, s^T)^{\mathrm{T}}1 = c,$$

where $\|1/n - r\|_2 \le \epsilon_2/(4\|AU\|_\infty^2)$ and $\|1/m - c\|_1 \le \epsilon_2/(4\|AU\|_\infty^2)$. Then we find that

$$F(u^T, v^T, s^T) = \min_\pi \left\{ \sum_{i,j} \pi_{i,j} M_{i,j} - \eta H(\pi) : \sum_j \pi_{i,j} = r_i, \sum_i \pi_{i,j} = c_j \right\},$$

and

$$\min_{u,v} F(u, v, s^T) = \min_\pi \left\{ \sum_{i,j} \pi_{i,j} M_{i,j} - \eta H(\pi) : \sum_j \pi_{i,j} = \frac{1}{n}, \sum_i \pi_{i,j} = \frac{1}{m} \right\},$$

where $M_{i,j} = \|A_{i,j} U s^T\|_2^2$. It follows that

$$\begin{aligned} &F(u^T, v^T, s^T) - \min_{u,v} F(u, v, s^T) \\ \le &\eta \log(mn) + 2\|1/m - c_1\|_1 \times \|AU\|_\infty^2 \le \epsilon_2, \end{aligned}$$

where the last inequality is by taking $\eta = \epsilon_2/(2\log(mn))$.

□

Lemma 9 (Lower Boundedness of F)**.** *Denote by (u^*, v^*, s^*) the global optimum of* (5). *Then we have that*

$$F(u^*, v^*, s^*) \ge 1 - \frac{1}{\eta}\|AU\|_\infty^2.$$

Proof of Lemma 9. It is easy to show that

$$\sum_{i,j} \pi_{i,j}(u^*, v^*, s^*) = 1.$$

Moreover, for any (i, j), we have that $c_{i,j} \le \|AU\|_\infty^2$. It follows that

$$\exp\left(-\frac{1}{\eta}\|AU\|_\infty^2 + u_i^* + v_j^*\right) \le \pi_{i,j} \le 1,$$

and therefore $u_i^* + v_j^* \le \frac{1}{\eta}\|AU\|_\infty^2$ for any (i, j). Hence we conclude that

$$\sum_{i,j} \pi_{i,j}(u^*, v^*, s^*) - \frac{1}{n}\sum_{i=1}^n u_i - \frac{1}{m}\sum_{j=1}^m v_j \ge 1 - \frac{1}{\eta}\|AU\|_\infty^2.$$

□

In the following we give a re-statement of Theorem 2 and the formal proof.

Theorem (Re-statement of Theorem 2)**.** *Choose parameters*

$$\tau = \frac{1}{4\|AU\|_\infty^2 L_2/\eta + \varrho L_1^2}, \quad \eta = \frac{\epsilon_2}{2\log(mn)}, \quad \varrho = \frac{2\|AU\|_\infty^2}{\eta} + \frac{4\|AU\|_\infty^4}{\eta^2},$$

and Algorithm 2 terminates when

$$\|\xi^{t+1}\| \le \epsilon_1, \quad \|1/n - \pi(u^t, v^t, s^t)1\|_2 \le \frac{\epsilon_2}{4\|AU\|_\infty^2}, \quad \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1 \le \frac{\epsilon_2}{4\|AU\|_\infty^2}.$$

We say that $(\hat{u}, \hat{v}, \hat{s})$ is a (ϵ_1, ϵ_2)-stationary point of (5) if

$$\|Grad_s F(\hat{u}, \hat{v}, \hat{s})\| \le \epsilon_1,$$
$$F(\hat{u}, \hat{v}, \hat{s}) - \min_{u,v} F(u, v, \hat{s}) \le \epsilon_2,$$

where $Grad_s F(u, v, s)$ denotes the partial derivative of F with respect to the variable s on the sphere $\mathbb{S}^{d(n+m)-1}$. Then Algorithm 2 returns an (ϵ_1, ϵ_2)-stationary point in iterations

$$T = \mathcal{O}\left(\log(mn) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2\epsilon_2}\right]\right).$$

Proof of Theorem 2. We can build the one-iteration descent result based on Lemma 5, Lemma 6, and Lemma 7:

$$\begin{aligned}
&F(u^{t+1}, v^{t+1}, s^{t+1}) - F(u^t, v^t, s^t) \\
\le& -\left(\frac{1}{2}\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \frac{1}{2}\|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \frac{1}{8\|AU\|_\infty^2 L_2/\eta + 2\varrho L_1^2}\|\xi^{t+1}\|_2^2\right) \\
=& -\frac{1}{2}\left(\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \frac{\eta^2\|\zeta^{t+1}\|^2}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right)
\end{aligned}$$

Then we have that

$$\begin{aligned}
&F(u^T, v^T, s^T) - F(u^0, v^0, s^0) \\
\le& -\frac{1}{2}\sum_{t=0}^{T-1}\left(\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \frac{\eta^2\|\zeta^{t+1}\|^2}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right) \\
\le& -\frac{1}{2}\cdot\min\left\{1, \frac{1}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right\} \\
&\qquad \times \sum_{t=0}^{T-1}\left(\|1/n - \pi(u^t, v^t, s^t)1\|_2^2 + \|1/m - \pi(u^{t+1}, v^t, s^t)^{\mathrm{T}}1\|_1^2 + \eta^2\|\xi^{t+1}\|_2^2\right) \\
\le& -\frac{1}{2}T\cdot\min\left\{1, \frac{1}{2\|AU\|_\infty^2\eta(2L_2 + L_1^2) + 4\|AU\|_\infty^4 L_1^2}\right\}\cdot\min\left\{\epsilon_1^2, \frac{\epsilon_2^2}{16\|AU\|_\infty^4}, \frac{\epsilon_2^2}{16\|AU\|_\infty^4}\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
T \leq & [F(u^0, v^0, t^0) - F(u^T, v^T, s^T)] \max\left\{2, 4\|AU\|_\infty^2 \eta(2L_2 + L_1^2) + 8\|AU\|_\infty^4 L_1^2\right\} \\
& \max\left\{\frac{1}{\epsilon_1^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}\right\} \\
\leq & \left(F(u^0, v^0, t^0) - 1 + \frac{\|AU\|_\infty^2}{\eta}\right) \max\left\{2, 4\|AU\|_\infty^2 \eta(2L_2 + L_1^2) + 8\|AU\|_\infty^4 L_1^2\right\} \\
& \max\left\{\frac{1}{\epsilon_1^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}, \frac{16\|AU\|_\infty^4}{\epsilon_2^2}\right\} \\
= & \mathcal{O}\left(\log(mn) \cdot \left[\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1^2 \epsilon_2}\right]\right).
\end{aligned}$$

□

E TECHNICAL PROOFS IN SECTION 4

E.1 Proof of Theorem 3

Proof of Lemma 1. Denote $\mathcal{F} = \{f \in \mathcal{H} : \ \|f\|_{\mathcal{H}} \leq 1\}$. By the bias-variation decomposition, we have that

$$\mathbb{E}[(\mathcal{KP}W(\hat{\mu}_n, \mu))^{1/p}] \leq \sup_{f\in\mathcal{F}} \ \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}] \\ + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left((W(f\#\hat{\mu}_n, f\#\mu))^{1/p} - \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}]\right)\right].$$

For fixed $f \in \mathcal{F}$, we can see that

$$\mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}] \leq c_p n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p}$$

where c_p is a constant depending only on p and

$$\zeta_{p,d} = \begin{cases} 1, & \text{if } d = 2p, \\ 0, & \text{otherwise.} \end{cases}$$

Now we start to upper bound the variation term. Define the empirical process

$$X_f = (W(f\#\hat{\mu}_n, f\#\mu))^{1/p} - \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}].$$

It is easy to see that $\mathbb{E}[X_f] = 0$. Moreover, we can show that for fixed f, the random variable X_f is sub-exponential. Denote by $Z = \{z_i\}_{i=1}^n$ and $Z' = \{z_i'\}_{i=1}^n$ i.i.d. samples from $f\#\mu$. Take $g(Z) = (W(f\#\hat{\mu}_n, f\#\mu))^{1/p}$. Then we have that

$$|g(Z) - g(Z'_{(i)})| \leq (W(f\#\hat{\mu}_n, f\#\hat{\mu}'_n))^{1/p} \leq n^{-1/(2\vee p)} \|Z - Z'\|_2.$$

It follows that

$$\sum_{i=1}^n \|\nabla_i g(Z)\|^2 \leq n^{-2/(2\vee p)}, \quad \max_{1\leq i\leq n} \|\nabla_i g(Z)\| \leq n^{-1/p}.$$

Then the Poincare's inequality in Theorem 7 implies that

$$\Pr\{X_f \geq t\} \leq \exp\left(-K^{-1} \min\{tn^{1/p}, t^2 n^{2/(2\vee p)}\}\right).$$

Hence we conclude that X_f is sub-exponential with parameters $(\sqrt{K/2} n^{-1/(2\vee p)}, (K/2) n^{-1/p})$.

For the function space $\mathcal{F}$, define the corresponding metric

$$\mathsf{d}(f, f') = \|f - f'\|_{\mathcal{H}}.$$

Let $X \sim \mu$. Then for any $f, f' \in \mathcal{F}$, we have that

$$\begin{aligned} &|X_f - X_{f'}| \\ \leq &\mathbb{E}\big[\, (W(f\#\hat{\mu}_n, f'\#\hat{\mu}_n))^{1/p} + (W(f\#\mu, f'\#\mu))^{1/p} \,\big] + \mathbb{E}\big[\, (W(f\#\hat{\mu}_n, f'\#\hat{\mu}_n))^{1/p} + (W(f\#\mu, f'\#\mu))^{1/p} \,\big] \\ \leq &2\left(\mathbb{E}\|f(X) - f'(X)\|_2^p\right)^{1/p} + \left(\frac{1}{n}\sum_{i=1}^n \|f(X_i) - f'(X_i)\|_2^p\right)^{1/p} + \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \|f(X_i) - f'(X_i)\|_2^p\right)^{1/p}\right]. \end{aligned}$$

Note that the following upper bound holds for any $f, f' \in \mathcal{F}$ and $x \in \mathbb{R}^D$:

$$\begin{aligned}
\|f(x) - f'(x)\|_2 &= \max_{a:\ \|a\|_2 \le 1} \langle f(x) - f'(x), a\rangle \\
&= \max_{a:\ \|a\|_2 \le 1} \langle f(x), a\rangle - \langle f'(x), a\rangle \\
&= \max_{a:\ \|a\|_2 \le 1} \langle f, K_x a\rangle_{\mathcal{H}_K} - \langle f', K_x a\rangle_{\mathcal{H}_K} \\
&= \max_{a:\ \|a\|_2 \le 1} \langle f - f', K_x a\rangle_{\mathcal{H}_K} \\
&\le \|f - f'\|_{\mathcal{H}_K} \times \max_{a:\ \|a\|_2 \le 1} \|K_x a\|_{\mathcal{H}_K} \\
&= \|f - f'\|_{\mathcal{H}_K} \times \max_{a:\ \|a\|_2 \le 1} \sqrt{a^{\mathrm{T}} K(x,x) a} \\
&= \sqrt{B}\|f - f'\|_{\mathcal{H}_K}.
\end{aligned}$$

As a consequence, substituting this upper bound into the relation above implies that

$$|X_f - X_{f'}| \le 4\sqrt{B}\mathsf{d}(f, f').$$

Applying the ϵ-net argument similar to the Dudley's entropy integral bound (Wainwright, 2019, Theorem 5.22) gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \le \inf_{\epsilon>0}\left\{4\sqrt{B}\epsilon + \sqrt{2K}n^{-1/(2\vee p)}\sqrt{\log \mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon)} + (K/2)n^{-1/p}\log \mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon)\right\}$$

Taking $\mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon) = \lceil \frac{1}{\epsilon} \rceil$ and $\epsilon = n^{-1/p}$ implies that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \lesssim n^{-1/(2\vee p)}\sqrt{\log(n)} + n^{-1/p}\log(n).$$

□

Proof of Lemma 2. We start to upper bound the variance term

$$\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} - \mathbb{E}[\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}].$$

Denote by $X = \{x_i\}_{i=1}^n$ and $X' = \{x_i'\}_{i=1}^n$ i.i.d. samples from μ, and let $g(X) = \left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}$. Based on the triangular inequality, we find that

$$\begin{aligned}
|g(X) - g(X')| &\le n^{-1/p}\left(\sum_{i=1}^n \max_{f\in\mathcal{F}} \|f(x_i) - f(x_i')\|_2\right)^{1/p} \\
&\le n^{-1/p}\left(\sum_{i=1}^n L\|x_i - x_i'\|\right)^{1/p} \\
&\le n^{-1/(2\vee p)} L^{1/p}\|X - X'\|.
\end{aligned}$$

It follows that

$$\sum_{i=1}^n \|\nabla_i g(Z)\|^2 \le n^{-2/(2\vee p)}L^{2/p}, \quad \max_{1\le i\le n}\|\nabla_i g(Z)\| \le n^{-1/p}L^{1/p}.$$

Then the Poincare's inequality in Theorem 7 implies that

$$\Pr\{\left|\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} - \mathbb{E}[\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}]\right| \ge t\} \le \exp\left(-K^{-1}\min\{tn^{1/p}L^{-1/p}, t^2 n^{2/(2\vee p)}L^{-2/p}\}\right).$$

Substituting the right-hand-side with α completes the proof. □

Proof of Theorem 3. Based on the triangular inequality, we can see that

$$\left| \left(\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m)\right)^{1/p} - \left(\mathcal{KPW}(\mu, \nu)\right)^{1/p} \right| \leq \left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} + \left(\mathcal{KPW}(\hat{\nu}_m, \nu)\right)^{1/p}.$$

It suffices to upper bound $\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}$ and $\left(\mathcal{KPW}(\hat{\nu}_m, \nu)\right)^{1/p}$ separately. By the bias-variance decomposition,

$$\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} \leq \mathbb{E}[\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}] + \left(\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} - \mathbb{E}[\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p}] \right),$$

where the first term quantifies the bias for empirical estimation, and the second term quantifies the variance of estimation. The bias term can be upper bounded by applying Lemma 1, and the variance term can be upper bounded by applying Lemma 2. In summary, with probability at least $1-\alpha$, it holds that

$$\begin{aligned}\left(\mathcal{KPW}(\hat{\mu}_n, \mu)\right)^{1/p} \lesssim & \max\left\{ n^{-1/p} K \log(1/\alpha), n^{-1/(2\vee p)} \sqrt{K \log(1/\alpha)} \right\} L^{1/p} \\ & + n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p} + n^{-1/(2\vee p)} \sqrt{\log(n)} + n^{-1/p} \log(n).\end{aligned}$$

The upper bound for $\left(\mathcal{KPW}(\hat{\nu}_m, \nu)\right)^{1/p}$ can be proceeded similarly.

□

E.2 Testing Performance

Based on the finite-sample guarantee in Theorem 3, we are able to characterize the performance of the KPW test. To make the type-I error below than α, we reject the null hypothesis as long as the empirical statistic $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m) \geq \gamma_{m,n}$, where

$$\begin{aligned}\gamma_{m,n}^{1/p} \sim & \max\left\{ N^{-1/p} K \log(1/\alpha), N^{-1/(2\vee p)} \sqrt{K \log(1/\alpha)} \right\} L^{1/p} \\ & + N^{-\frac{1}{(2p)\vee d}} (\log N)^{\zeta_{p,d}/p} + N^{-1/(2\vee p)} \sqrt{\log(n)} + N^{-1/p} \log(n).\end{aligned}$$

For the alternative hypothesis, assume that target distributions μ and ν satisfy $\mathcal{KPW}(\mu, \nu) > \gamma_{m,n}$. Then the type-II error can be upper bounded as

$$\begin{aligned}& \Pr_{\mathcal{H}_1}\left(\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m) < \gamma_{m,n} \right) \\ =& \Pr_{\mathcal{H}_1}\left(\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m) - \mathcal{KPW}(\mu, \nu) < \gamma_{m,n} - \mathcal{KPW}(\mu, \nu) \right) \\ =& \Pr_{\mathcal{H}_1}\left(\mathcal{KPW}(\mu, \nu) - \mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m) > \mathcal{KPW}(\mu, \nu) - \gamma_{m,n} \right) \\ \leq& \Pr_{\mathcal{H}_1}\left(|\mathcal{KPW}(\mu, \nu) - \mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m)| > \mathcal{KPW}(\mu, \nu) - \gamma_{m,n} \right) \\ \leq& \frac{\mathbb{E}\left(\mathcal{KPW}(\mu, \nu) - \mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_m)\right)^2}{\left(\mathcal{KPW}(\mu, \nu) - \gamma_{m,n}\right)^2}.\end{aligned}$$

E.3 Finite-sample Guarantee for $p \in [1, 2)$

In this subsection, we discuss the finite-sample guarantee for KPW distance with p-Wasserstein distance for $p \in [1, 2)$. Note that it is not necessary to rely on the Poincare inequality or projection poincare inequality to obtain the result. We first present several technical lemmas before showing the final result.

Lemma 10. *Based on Assumption 1, for $f \in \{f \in \mathcal{H} : \ \|f\|_{\mathcal{H}} \leq 1\}$, we have*

$$\|f(x)\|_2 \leq \sqrt{B}, \quad \forall x \in \mathbb{R}^D.$$

Proof of Lemma 10. For fixed $x \in \mathcal{X}$, the norm of $f(x)$ can be upper bounded as the following:

$$\|f(x)\|_2^2 = \langle f(x), f(x) \rangle = \langle f, K_x f(x) \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|K_x f(x)\|_{\mathcal{H}} \leq \|K_x f(x)\|_{\mathcal{H}}.$$

In particular,

$$\begin{aligned}\|K_x f(x)\|_{\mathcal{H}}^2 &= \langle K_x f(x), K_x f(x)\rangle_{\mathcal{H}} \\ &= \langle \big(K_x f(x)\big) f(x), f(x)\rangle \\ &= \langle K(x, f(x)) f(x), f(x)\rangle \\ &= f(x)^{\mathrm{T}} K(x, f(x)) f(x) \\ &\leq B\|f(x)\|_2^2\end{aligned}$$

Combining those two relations above implies the desired result. □

Lemma 11. *For $p \in [1, 2)$, the bias term of empirical KPW distance can be upper bounded as*

$$\mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \lesssim n^{-\frac{1}{(2p)\vee d}} (\log n)^{\zeta_{p,d}/p} + n^{1/2-1/p}\sqrt{\log(n)} + n^{-1/p}.$$

where $\zeta_{p,d} = 1$ if $d = 2p$ and $\zeta_{p,d} = 0$ otherwise.

Proof of Lemma 11. Following the similar argument as in Lemma 1, we can see that

$$\begin{aligned}\mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}] \leq \sup_{f\in\mathcal{F}} \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}] \\ + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\Big((W(f\#\hat{\mu}_n, f\#\mu))^{1/p} - \mathbb{E}[(W(f\#\hat{\mu}_n, f\#\mu))^{1/p}]\Big)\right],\end{aligned}$$

and the first term can also be bounded similarly. To upper bound the second term, define the empirical process $\{X_f\}$ as in Lemma 1. For fixed f, the random variable X_f can be shown to be sub-Gaussian. Denote by $Z = \{z_i\}_{i=1}^n$ and $Z'_{(i)}$ the sample set so that the i-th element is different. Take $g(Z) = (W(f\#\hat{\mu}_n, f\#\mu))^{1/p}$. Then we have that

$$\begin{aligned}|g(Z) - g(Z'_{(i)})| \leq (W(f\#\hat{\mu}_n, f\#\hat{\mu}'_n))^{1/p} &\leq \left(\frac{1}{n}\|f(z_i) - f(z'_i)\|_2^p\right)^{1/p} \\ &\leq n^{-1/p} 2\sqrt{B}.\end{aligned}$$

Therefore, applying the McDiarmid's inequality in Theorem 6 implies

$$\Pr\{|X_f| \geq u\} \leq 2\exp\left(-\frac{u^2}{2Bn^{1-2/p}}\right).$$

Applying Lemma 3 implies that for fixed ℓ, the random variable X_f is sub-Gaussian with the parameter $\sigma^2 = 36Bn^{1-2/p}$. Then applying the ϵ-net argument similar to the Dudley's entropy integral bound (Wainwright, 2019, Theorem 5.22) gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \leq \inf_{\epsilon>0}\left\{4\sqrt{B}\epsilon + \sqrt{36Bn^{1-2/p}}\sqrt{2\log\mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon)}\right\}.$$

Taking $\mathcal{N}(\mathcal{F}, \mathsf{d}, \epsilon) = \lceil \frac{1}{\epsilon} \rceil$ and $\epsilon = n^{-1/p}$ implies that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \lesssim n^{1/2-1/p}\sqrt{\log(n)} + n^{-1/p}.$$

□

Lemma 12. *For $p \in [1, 2)$, with with probability at least $1 - \alpha$, it holds that*

$$\left|(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}]\right| \leq n^{1/2-1/p}\sqrt{2B\log\frac{2}{\alpha}}.$$

Proof of Lemma 12. Denote by $Z = \{z_i\}_{i=1}^n$ and $Z'_{(i)}$ the sample set so that the i-th element is different. Take $g(Z) = (\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}$. Then we can see that

$$|g(Z) - g(Z'_{(i)})| \leq (\mathcal{KPW}(\hat{\mu}_n, \hat{\mu}'_n))^{1/p} \leq n^{-1/p} 2\sqrt{B}.$$

Then applying the McDiarmid's inequality in Theorem 6 implies

$$\Pr\left\{\left|(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p} - \mathbb{E}[(\mathcal{KPW}(\hat{\mu}_n, \mu))^{1/p}]\right| \geq u\right\} \leq 2\exp\left(-\frac{u^2}{2Bn^{1-2/p}}\right).$$

□

Based on Lemma 11 and Lemma 12, we obtain the uncertainty quantification result in Theorem 4.

F IMPLEMENTATION DETAILS FOR COMPUTING KPW DISTANCE

The variable s is initialized to be a uniform random vector over sphere. The dual variable v is initialized to be a Gaussian random vector with unit covariance. When updating the block of variables u^{t+1} and v^{t+1}, we make the change of variables $(u')^{t+1} = \exp(u^{t+1})$ and $(v')^{t+1} = \exp(v^{t+1})$. We update $(u')^{t+1}$ and $(v')^{t+1}$ instead to accelerate the computation:

$$(u')^{t+1} = \left\{ \frac{1/n}{\sum_j \exp\left(-\frac{1}{\eta}c_{i,j} + (v'_j)^t\right)} \right\}_i$$
$$(v')^{t+1} = \left\{ \frac{1/m}{\sum_i \exp\left(-\frac{1}{\eta}c_{i,j} + (u'_i)^{t+1}\right)} \right\}_j,$$

and we further store the matrix A with $A_{i,j} = \exp\left(-\frac{1}{\eta}c_{i,j}\right)$ in advance to reduce the computational cost. The transport mapping $\pi^{t+1} \triangleq (\pi_{i,j}(u^{t+1}, v^{t+1}, s^t))_{i,j}$ can be formulated without going through a for loop but only with multiplication operators:

$$\pi^{t+1} = (u')^{t+1} \ \texttt{.*}\ A \ \texttt{.*}\ [(v')^{t+1}]^{\mathrm{T}},$$

where the operator `.*` means we multiply two objects componentiwisely in terms of array broadcasting. When updating ζ^{t+1}, we first formulate the matrix V^{t+1} with

$$V^{t+1}_{i,j} = \sum_{i,j} \pi^{t+1}_{i,j} A^{\mathrm{T}}_{i,j} A_{i,j}$$

and then continue the matrix multiplication procedure in (6i). Denote by G_i the i-th row block of the gram matrix G, then

$$V^{t+1} = \left\{ \sum_{i,j} \pi^{t+1}_{i,j} (G_i + G_{n+j})^{\mathrm{T}} (G_i + G_{n+j}) \right\}_{i,j}$$
$$= \left\{ \sum_{i,j} \pi^{t+1}_{i,j} (G_i^{\mathrm{T}} G_i + G_{n+j}^{\mathrm{T}} G_{n+j} + G_{n+j}^{\mathrm{T}} G_i + G_i^{\mathrm{T}} G_{n+j}) \right\}_{i,j}.$$

Consequently, we can compute each of the four components in the formula above without executing double for loops and then sum them up to obtain the matrix V^{t+1}. During the numerical implementation, we also find that the computation is sensitive to the choice of η. This phenomenon has also been observed when using Sinkhorn's algorithm to compute Wasserstein distance or projected Wasserstein distance. When η is too small, the iteration update may have numerical instability issues. When η is too large, the obtained solution is far away from the optimal solution to the original KPW distance. We have tried the best to tune this parameter to make the algorithm maintain the best performance. How to tune this hyper-parameter systematically is left for future works.

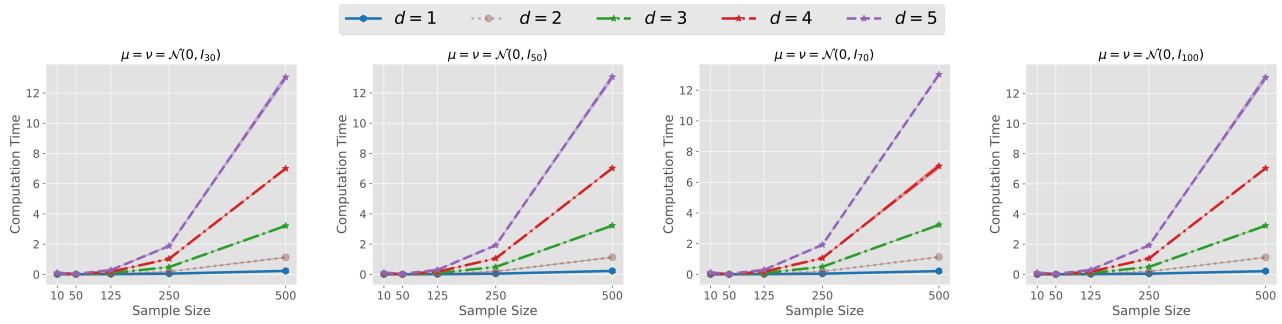

Figure 4: Mean computation time for computing $\mathcal{KPW}(\hat{\mu}_n, \hat{\nu}_n)$ for varying n. Results are averaged over 10 independent trials.

G DETAILS ABOUT EXPERIMENT

G.1 Sample Complexity

In this experiment, we fix hyper-parameters $\sigma^2 = 1, \rho = 0.5$ for computing KPW distances. The values of empirical KPW distances across different choices of sample size are reported in Figure 1, and the corresponding computation time is reported in Figure 4. From the plot we can see that it is efficient to compute KPW distances with reasonably small sample size n and projected dimension d.

G.2 Configurations

All methods are implemented using python 3.7 (Pytorch 1.1) on a MacBook Pro labtop with 32GB of memory. When running the code, there is no swapping of memory and the average CPU frequency is 3.2 GHZ. We compute the projected Wasserstein distance based on the official code in `https://github.com/fanchenyou/PRW`. We run the MMD-O test based on the code in `https://github.com/fengliu90/DK-for-TST`. We run the MMD-NTK test based on the code in `https://github.com/xycheng/NTK-MMD`. From extensive experiments we realize that MMD-NTK is the most computationally efficient test, but its power does not scale the best. On the other hand, this method can be useful when performing a test for the large-sampled case, while our method may be intractable to compute in short time. We run the ME test based on the code in `https://github.com/wittawatj/interpretable-test`.

G.3 Implementation of Cross-Validation

The candidate choices of hyper-parameters ρ and σ^2 are within the set

$$\{(\rho, \sigma^2) : \ \sigma^2 = a \cdot \hat{\sigma}^2 : \ a \in \{0.5, 1, 2\}, \rho \in \{0.25, 0.5, 0.75\}\},$$

where $\hat{\sigma}^2$ denotes the empirical median of pairwise distances between observations. To choose ρ and σ^2, we further split the training set into the training and validation dataset, which contain 70% and 30% data, respectively. For each choice of hyper-parameters we use the training dataset to obtain a nonlinear projector and examine its hold-out performance on the validation dataset, which is quantified as the negative of the p-value for two-sample tests between two collection of samples in the validation dataset. We choose hyper-parameters ρ and σ^2 with the best hold-out performance.

G.4 Tests for Synthetic Datasets

When studying tests on Gaussian distributions, we take both the training and testing sample sizes N to be 50. When reproducing the experiments corresponding to the left two figures in Fig. 3, we take the dimension $D \in \{20, 40, 60, 80, 100, 120, 140, 160\}$. When reproducing the experiments corresponding to the right two figures, we take the sample size $n = m \in \{80, 100, 140, 180, 250\}$.

Table 3: Average type-I error and standard error for two-sample tests in *MNIST* dataset across different choices of sample size.

N	MMD-NTK	MMD-O	ME	PW	KPW
200	0.057±0.0010	0.056±0.0006	0.044±0.0003	0.056±0.0004	0.061±0.0005
250	0.051±0.0003	0.060±0.0001	0.065±0.0002	0.046±0.0003	0.048±0.0002
300	0.068±0.0006	0.055±0.0003	0.059±0.0007	0.056±0.0002	0.053±0.0001
400	0.049±0.0007	0.058±0.0002	0.041±0.0002	0.061±0.0006	0.056±0.0006
500	0.061±0.0006	0.054±0.0004	0.060±0.0002	0.049±0.0003	0.047±0.0004
Avg.	0.057	0.056	0.053	0.054	0.053

G.5 Tests for MNIST handwritten digits

Table 3 present the type-I error for various tests in MNIST dataset, from which we can see that all tests have the type-I error close to $\alpha = 0.05$.

G.6 Human activity detection

The pre-processing of data is as follows. We first remove frames in which the person is standing still or with little movements. Then we delete the first few frames to make the action of bending consist of 500 frames. Next we delete the last few frames to make the action of throwing consist of 355 frames. We take the window size $W = 100$. To perform online change point detection, we pre-train a nonlinear projector using the data before time index 300 and compute the null statistics for many times to obtain the true threshold. Then we compute the detection statistic by comparing the distribution between the block of data before time 300 and the data from the sliding window. We reject the null hypothesis and claim a change is happened if the statistic is above the threshold. The plot of the detection statistic over time after the time index 400 is presented in Fig. 5, and the delay detection time corresponding to all users are reported in Table 2.

H IMPACT OF HYPER-PARAMETERS

H.1 Impact of Projected Dimension d

We prefer to choose the projected dimension d with relatively small values since the testing statistic will have poor sample complexity rate and is expansive to compute for large d. In this section, we examine the testing performance for different choices of d. In particular, we perform the KPW test on Gaussian distributions (with diagonal covariance matrices, $D = 128$ and $n = m = 50$) and Gaussian mixture distributions (with $D = 100$ and $n = m = 100$) following the setup in Section 5.1, the results of which are reported in Fig. 6. From the plot we can see that the testing power is generally better for $d > 1$, which suggests that using vector-valued RKHS is better than using classical scalar-valued RKHS. Moreover, we observe the performance is insensitive to the choice of d as long as we take $d > 1$.

H.2 Impact of Entropic Regularization Parameter η

As pointed out in Genevay et al. (2019), the entropic regularization in (4) could alerady improve the sample complexity result of Wasserstein distance. We perform experiments in this subsection to validate the impact of the entropic regularization parameter η for the performance of KPW test. The generated data follows Gaussian distributions (with $n = m = 100$) or Gaussian mixture distributions (with $n = m = 200$) with different choices of dimension D and fixed sample size. Benchmark methods include 1) *KPW test with* $\eta = 0$ (here Wasserstein distance is computed exactly and we apply alternating optimization procedure as a heuristic); 2) *Sinkhorn test* with the same η as in the KPW test (in which we take the Sinkhorn divergence as the statistic and all training and testing samples are used); 3) *Sinkhorn+* (using all data and post-selecting η with the best performance). Experiment results are reported in Fig. 7, from which we can see that even Sinkhorn+ test has the curse of dimension issue. Moreover, the KPW test with $\eta = 0$ has similar performance as the KPW test. Hence, we can assert that the KPW test is capable of alleviating the curse of dimension mainly due to the kernel projection operator instead of the entropic regularization.

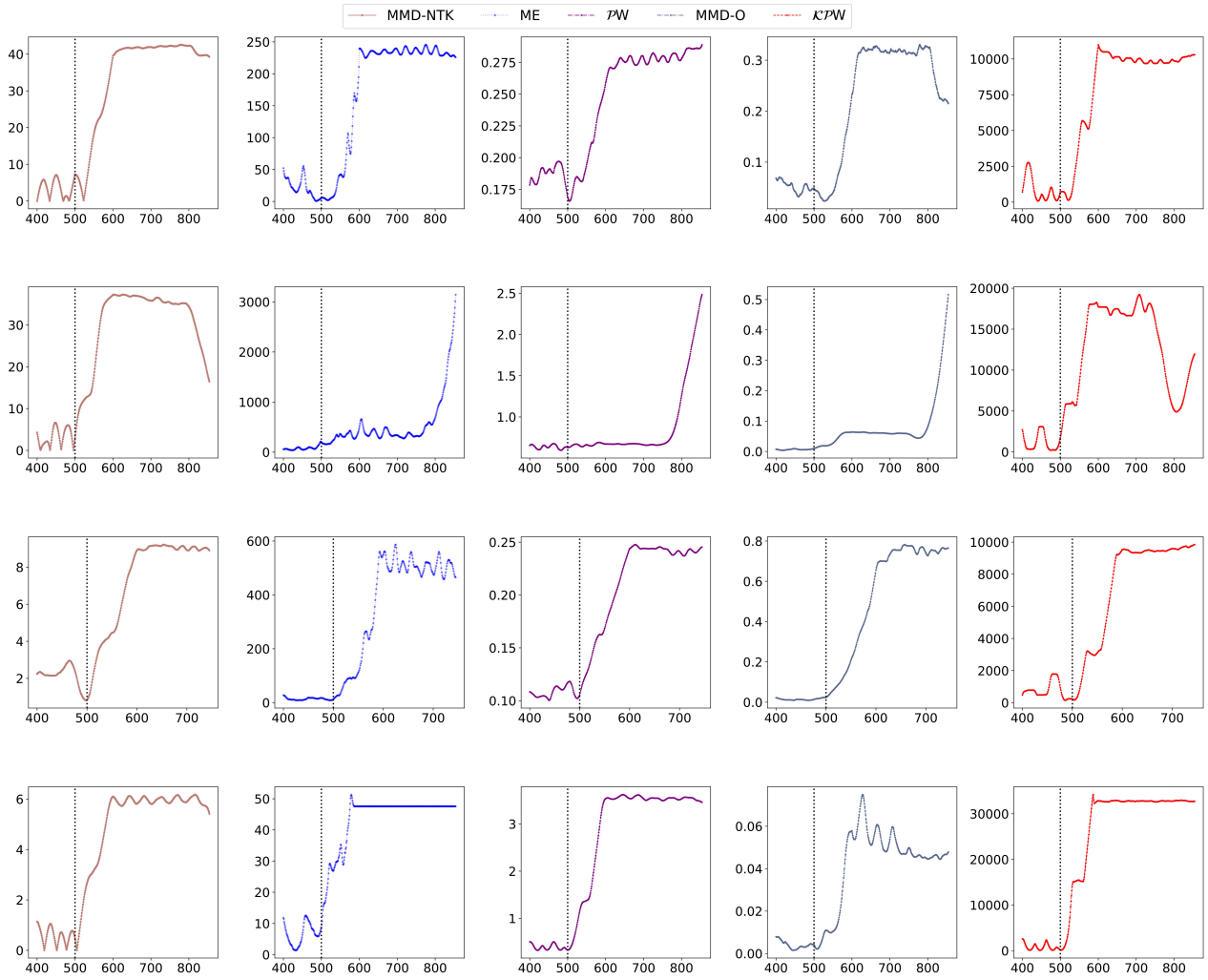

Figure 5: Comparison of detection statistics from bending to throwing for various testing procedures. Black dash line indicates the true change-point. Each row corresponds to detection results for each user.

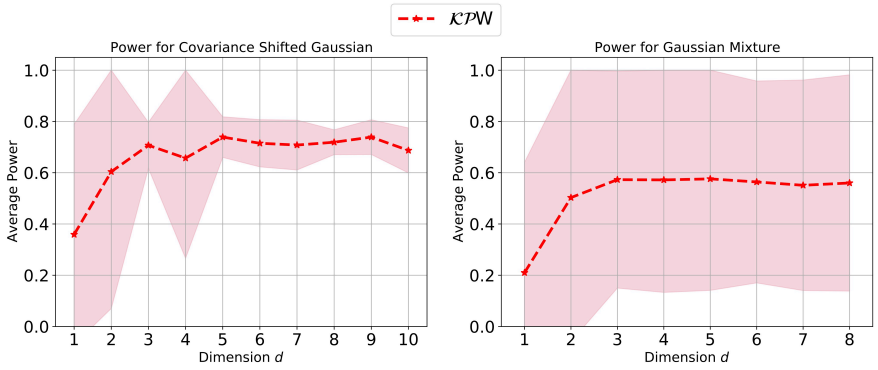

Figure 6: Average power for KPW test across different choices of projected dimension d. Left: Gaussian distribution; Right: Gaussian mixture distribution. Results are averaged over 10 independent trials.

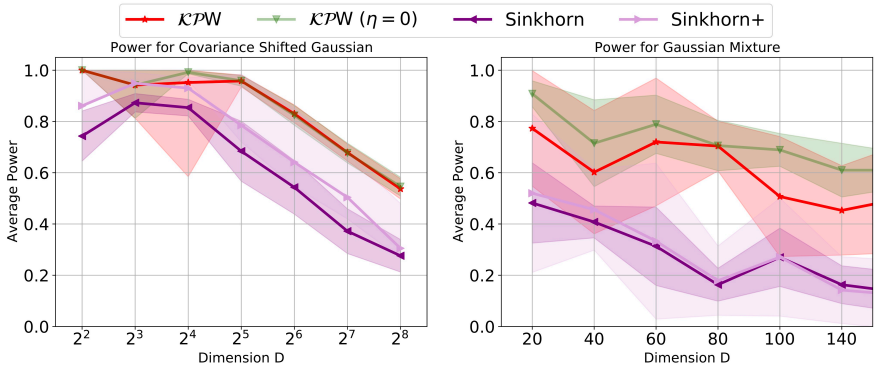

Figure 7: Average power for KPW tests and Sinkhorn tests across different choices of data dimension D. Left: Gaussian distribution; Right: Gaussian mixture distribution. Results are averaged over 10 independent trials.

I SOCIETAL IMPACT

Two-sample testing is not only a fundamental problem in statistics but also growing increasing attention in machine learning. On the one hand, it plays a key role in modern applications such as anomaly detection and health care. On the other hand, it can help to design better algorithms for artificial intelligence such as GANs. Our work shows a competitive performance for dealing with high-dimensional data by nonlinear dimensionality reduction using kernel trick. It identifies the difference between two collections of samples by extracting the most representative nonlinear features. We hope this work can be applied to design more powerful algorithms in those areas.