Statistical and Computational Guarantees of Kernel Max-Sliced Wasserstein Distances

Jie Wang
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, jwang3163@gatech.edu

March Boedihardjo
Department of Mathematics, Michigan State University, East Lansing, MI 48824, boedihar@msu.edu

Yao Xie
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, yao.xie@isye.gatech.edu

Optimal transport has been very successful for various machine learning tasks; however, it is known to suffer from the curse of dimensionality. Hence, dimensionality reduction is desirable when applied to high-dimensional data with low-dimensional structures. The kernel max-sliced Wasserstein distance is developed for this purpose by finding an optimal nonlinear mapping that reduces data into 1 dimensions before computing the Wasserstein distance. However, its theoretical properties have not yet been fully developed. In this paper, we provide sharp finite-sample guarantees under milder technical assumptions compared with state-of-the-art for the kernel projected p-Wasserstein distance between two empirical distributions with n samples for general $p \in [1, \infty)$. Algorithm-wise, we show that computing the kernel projected 2-Wasserstein distance is NP-hard, and then we further propose a semidefinite relaxation (SDR) formulation (which can be solved efficiently in polynomial time) and provide a relaxation gap for the SDP solution. We provide numerical examples to demonstrate the good performance of our scheme for high-dimensional two-sample testing.

1. Introduction

Optimal transport has achieved much success in various areas, such as generative modeling [25, 56, 45, 54], distributional robust optimization [23, 24, 68], non-parametric testing [74, 58, 73, 71, 66], domain adaptation [2, 14, 12, 13, 72], etc. See [57] for comprehensive reviews on these topics.

When applying optimal transport (OT) in statistical inference, one usually cares about the sample complexity of Wasserstein distance, i.e., how close between a population distribution μ and its empirical distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ with $x_i \sim \mu$ in terms of the "Wasserstein distance". Unfortunately, one needs the sample size n to be exponentially large in data dimension to achieve accurate enough estimation [21], referred to as the *curse of dimensionality* issue.

To tackle challenge of high dimensionality, it is meaningful to combine OT with projection operators to low-dimensional spaces. Researchers first attempted to study Sliced Wasserstein distances [10, 11, 17, 37, 36, 49, 51], which compute the average of the Wasserstein distance between two projected distributions using random one-dimensional projections. Since a single random projection contains little information to distinguish two high-dimensional distributions, computing the sliced Wasserstein distance requires a large number of linear projections. To tackle this issue, more recent literature considered the **M**ax-**S**liced (**MS**) Wasserstein distance that seeks the *optimal* projection direction such that the Wasserstein distance between projected distributions is maximized [16, 41, 43, 55, 69]. Later Wang et al. [70] modified the max-sliced Wasserstein distance by seeking an optimal nonlinear projection belonging to a ball of reproducing kernel Hilbert space (RKHS), which we call the **K***ernel* **M***ax-***S***liced (***KMS***) Wasserstein distance*. The motivation is that a nonlinear projector can be more flexible in capturing the differences between two high-dimensional distributions; it is worth noting that KMS Wasserstein reduces to MS Wasserstein when specifying a dot product kernel.

Despite promising applications of KMS Wasserstein distance, its computational and statistical results are still limited. From the computational perspective, Wang et al. [70] designed a gradient-based

algorithm to find local optimal points for computing the empirical KMS Wasserstein distance, which is inspired by the efficient manifold gradient algorithm in [31]. However, there is no theoretical guarantee regarding the quality of the obtained local optimum solution. In numerical experiments, the quality of the local optimum solution is highly sensitive to the choice of the initial iteration point. From the statistical perspective, the authors therein built concentration properties of the empirical KMS Wasserstein distance for distribution satisfying the projection Poincare inequality and Poincare inequality, which could be difficult to verify in practice.

To improve the aforementioned limitations, in this paper, we provide new computational and statistical results regarding the KMS Wasserstein distance. The following summarizes our contributions.

Sharp Finite-Sample Guarantees of KMS p-Wasserstein. We provide a non-asymptotic estimate on the KMS p-Wasserstein distance between two empirical distributions based on n samples, referred to as the *finite-sample guarantees*. Our result shows that when the samples are drawn from identical populations, the rate of convergence is $n^{-1/(2p)}$, which is dimension-free and optimal in the worst case.

Computation of KMS 2-Wasserstein. We analyze the computation of KMS 2-Wasserstein distance between two empirical distributions based on n samples. First, we show that computing this distance exactly is NP-hard. The proof methodology involves reducing the NP-hard fair-PCA problem, which focuses on maximizing the minimization of homogeneous quadratic functions [64], to this specific problem. Consequently, we are prompted to propose a semidefinite relaxation (SDR) as an approximate heuristic.

We further propose an efficient first-order method with biased gradient oracles to solve the SDR formulation, the complexity of which for finding a δ-optimal solution is $\widetilde{\mathcal{O}}\left(n^2\delta^{-3}\right)$. In comparison, the complexity of the interior point method for solving SDR is $\widetilde{\mathcal{O}}(n^{6.5})$. Next, we derive theoretical guarantees regarding the optimal solutions from SDR. We show that there exists an optimal solution from SDR that is at most rank-k, where $k \triangleq 1 + \lfloor \sqrt{2n + 9/4} - 3/2 \rfloor$, whereas computing the KMS distance exactly requires a rank-1 solution. An intuitive explanation is that we show that any extreme point of SDR is at most rank-k, and the set of extreme points of SDR must have a non-empty intersection with the set of its optimal solutions. We also provide a corresponding rank reduction algorithm designed to identify such low-rank solutions from the pool of optimal solutions of SDR.

Numerical Studies. Finally, we exemplify our theoretical results in two numerical studies: the uncertainty quantification and non-parametric two-sample testing. Our numerical results showcase the stable performance and quick computational time of our SDR formulation, as well as the desired sample complexity rate of the empirical KMS Wasserstein distance.

Literature. The study on the statistical and computational results of MS and KMS Wasserstein distances is popular in the existing literature. From the statistical perspective, existing results on the rate of empirical MS/KMS Wasserstein are either dimension-dependent, suboptimal or require regularity assumptions (e.g., log-concavity, Poincare inequality, projection Bernstein tail condition) on the population distributions [52, 3, 43, 69], except the very recent literature [9] that provides a sharp, dimension-free rate for MS Wasserstein with data distributions supported on a compact subspace but without regularity assumptions. From the computational perspective, there are two main approaches to compute such distances. One is to apply gradient-based algorithms to find local optimal solutions or stationary points, see, e.g., [42, 32, 30, 33, 70] Unfortunately, due to the highly non-convex nature of the optimization problem, the quality of the estimated solution is unstable and highly depends on the choice of initial guess. The other is to consider solving its SDR instead [55], yet the theoretical guarantees on the solution from convex relaxation are missing. Inspired from existing reference [4, 18, 53, 40] that studied the rank bound of SDR for various applications, we adopt their proof techniques to provide similar guarantees for computing KMS in Theorem 5. Besides, all listed references add entropic regularization to the inner optimal transport problem and solve the regularized version of MS/KMS Wasserstein distances instead, while the gap between the solutions from regularized and original problems could be non-negligible.

2. Background

We first introduce the definition of Wasserstein and KMS Wasserstein distances below.

DEFINITION 1 (WASSERSTEIN DISTANCE). Let $p \in [1, \infty)$. Given a normed space $(\mathcal{V}, \|\cdot\|)$, the p-Wasserstein distance between two probability measures μ, ν on $\mathcal{V}$ is defined as

$$W_p(\mu, \nu) = \left(\min_{\pi \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, \mathrm{d}\pi(x, y) \right)^{1/p}$$

where $\Gamma(\mu, \nu)$ denotes the set of all probability measures on $\mathcal{V} \times \mathcal{V}$ with marginal distributions being μ and ν.

DEFINITION 2 (REPRODUCING KERNEL HILBERT SPACE (RKHS)). Consider a symmetric and positive definite kernel $K : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$, where $\mathcal{B} \subseteq \mathbb{R}^d$. Given such a kernel, there exists a unique Hilbert space $\mathcal{H}$, called the RKHS, associated with the reproducing kernel K. Denote by K_x the kernel section at $x \in \mathcal{B}$ defined by $K_x(y) = K(x, y), \forall y \in \mathcal{B}$. Any function $f \in \mathcal{H}$ satisfies the reproducing property $f(x) = \langle f, K_x \rangle_{\mathcal{H}}, \forall x \in \mathcal{B}$. For $x, y \in \mathcal{B}$, it holds that $K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}}$.

DEFINITION 3 (KERNEL MAX-SLICED (KMS) WASSERSTEIN DISTANCE). Let $p \in [1, \infty)$. Given two distributions μ and ν, define the p-KMS Wasserstein distance as

$$\mathcal{KMS}_p(\mu, \nu) = \max_{f \in \mathcal{H},\ \|f\|_{\mathcal{H}} \le 1} W_p(f_{\#}\mu, f_{\#}\nu),$$

where $f_{\#}\mu$ and $f_{\#}\nu$ are the pushforward measures of μ and ν by $f : \mathcal{B} \to \mathbb{R}$, respectively.

In particular, for dot product kernel $K(x, y) = x^{\mathrm{T}} y$, the RKHS $\mathcal{H} = \{f :\ f(x) = x^{\mathrm{T}}\beta, \exists \beta \in \mathbb{R}^d\}$. In this case, the KMS Wasserstein distance reduces to the max-sliced Wasserstein distance [16]. A more flexible choice is the Gaussian kernel $K(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|_2^2)$, where $\sigma > 0$ denotes the temperature hyper-parameter. In this case, the function class $\mathcal{H}$ satisfies the *universal property* as it is dense in the continuous function class.

Given the RKHS $\mathcal{H}$, let the *canonical feature map* that embeds data to $\mathcal{H}$ as

$$\Phi :\ \mathcal{B} \to \mathcal{H}, \qquad x \mapsto \Phi(x) = K_x. \tag{1}$$

Define the functional $u_f :\ \mathcal{H} \to \mathbb{R}$ by $u_f(g) = \langle f, g \rangle_{\mathcal{H}}$ for any $g \in \mathcal{H}$, which can be viewed as a linear projector that maps data from the Hilbert space $\mathcal{H}$ to the real line. In light of this, for two probability measures μ and ν on $\mathcal{H}$, we define the MS p-Wasserstein distance

$$\mathcal{MS}_p(\mu, \nu) = \sup_{f \in \mathcal{H}:\ \|f\|_{\mathcal{H}} \le 1} W_p\Big((u_f)_{\#}\mu, (u_f)_{\#}\nu\Big), \tag{2}$$

where $(u_f)_{\#}\mu$ denotes the pushforward measure of μ by the map u_f, i.e., if μ is the distribution of a random element X of $\mathcal{H}$, then $(u_f)_{\#}\mu$ is the distribution of the random variable $u_f(X) = \langle f, X \rangle$, and $(u_f)_{\#}\nu$ is defined likewise. In the following, we show that the KMS Wasserstein distance in Definition 3 can be reformulated as the MS Wasserstein distance between two distributions on (infinite-dimensional) Hilbert space.

REMARK 1 (REFORMULATION OF KMS WASSERSTEIN). By the reproducing property, we can see that $f(x) = \langle f, K_x \rangle_{\mathcal{H}} = u_f(\Phi(x))$, which implies $f = u_f \circ \Phi$. As a consequence,

$$\mathcal{KMS}_p(\mu, \nu) = \sup_{f \in \mathcal{H}:\ \|f\|_{\mathcal{H}} \le 1} W_p\Big((u_f)_{\#}\big(\Phi_{\#}\mu\big), (u_f)_{\#}\big(\Phi_{\#}\nu\big)\Big) = \mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\nu\Big). \tag{3}$$

In other words, the KMS Wasserstein distance first maps data points into the infinite-dimensional Hilbert space $\mathcal{H}$ through the canonical feature map Φ, and next finds the linear projector to maximally distinguish data from two populations. Compared with the traditional MS Wasserstein distance [16] that performs linear projection in $\mathbb{R}^d$, KMS Wasserstein distance is a more flexible notion.

REMARK 2 (CONNECTIONS WITH KERNEL PCA). Given data points $x_1, \ldots, x_n$ on $\mathcal{B}$, denote by $\widehat{\mu}_n$ the corresponding empirical distribution. Assume $\frac{1}{n}\sum_{i\in[n]} \Phi(x_i) = 0$, since otherwise one can center those data points as a preprocessing step. Kernel PCA [47] is a popular tool for nonlinear dimensionality reduction. When seeking the first principal nonlinear projection function f, [46] presents the following reformulation of kernel PCA:

$$\underset{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\le 1}{\arg\max}\ \mathrm{Var}\Big((u_f)_{\#}(\Phi_{\#}\widehat{\mu}_n)\Big), \tag{4}$$

where $\mathrm{Var}(\cdot)$ denotes the variance of a given probability measure. In comparison, the KMS Wasserstein distance aims to find the optimal nonlinear projection function that distinguishes two populations and replaces the variance objective in (4) with the Wasserstein distance between two projected distributions in (3). Also, kernel PCA is a special case of KMS Wasserstein by considering $p = 2, \mu \equiv \widehat{\mu}_n, \nu \equiv \delta_0$ in (3).

Notations. We use "MATLAB notation" to define block matrics: for matrices $A_1, \ldots, A_n$ of common width, let $[A_1, \ldots, A_n]$ denote the matrix obtained by horizontal contamination of them, and $[A_1; \ldots; A_n]$ denote the matrix obtained by vertical contamination of them. Let $\langle \cdot, \cdot \rangle$ denote the inner product operator. For any positive integer n, denote $[n] = \{1, 2, \ldots, n\}$. Define

$$\Gamma_n = \Big\{\pi \in \mathbb{R}_+^{n\times n} :\ \sum_{i=1}^n \pi_{i,j} = \frac{1}{n}, \sum_{j=1}^n \pi_{i,j} = \frac{1}{n}, \quad \forall i, j \in [n]\Big\}. \tag{5}$$

Let $\mathrm{Conv}(P)$ denote a convex hull of the set P, and $\mathbb{S}_n^+$ denote the set of positive semidefinite matrices of size $n \times n$. We use $\widetilde{\mathcal{O}}(\cdot)$ as a variant of $\mathcal{O}(\cdot)$ to hide logarithmic factors.

3. Statistical Guarantees

Suppose samples $x^n := \{x_i\}_{i\in[n]}$ and $y^n := \{y_i\}_{i\in[n]}$ are given and follow distributions μ, ν, respectively. Denote by $\widehat{\mu}_n$ and $\widehat{\nu}_n$ the corresponding empirical distributions from samples x^n and y^n. In this section, we provide a finite-sample guarantee on the p-KMS Wasserstein distance between $\widehat{\mu}_n$ and $\widehat{\nu}_n$ with $p \in [1, \infty)$. This guarantee can be helpful for KMS Wasserstein distance-based hypothesis testing: Suppose one aims to build a non-parametric test to distinguish two hypotheses $H_0 :\ \mu = \nu$ and $H_1 :\ \mu \neq \nu$. Thus, it is crucial to control the high-probability upper bound of $\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n)$ under H_0 as it serves as the critical value to determine whether H_0 is rejected or not. We first make the following assumption on the kernel.

ASSUMPTION 1. *The kernel $K(\cdot, \cdot)$ satisfies that $\sqrt{K(x, x)} \le A$ for any $x \in \mathcal{B}$.*

Assumption 1 is standard in the literature (see, e.g., [26]), and is quite mild: The Gaussian kernel $K(x, y) = \exp(-\|x - y\|_2^2/\sigma^2)$ naturally fits into this assumption. For dot product kernel $K(x, y) = x^{\mathrm{T}}y$, if we assume the support $\mathcal{B}$ has a finite diameter, this assumption can also be satisfied. Define the critical value

$$\Delta(n, \alpha) = 4A\Big(C + 4\sqrt{\log\frac{2}{\alpha}}\Big)^{1/p} \cdot n^{-1/(2p)},$$

where $C \ge 1$ is a universal constant. Now, we have the following finite-sample guarantees on KMS p-Wassersrein distance.

THEOREM 1 **(Finte-Sample Guarantee)**. *Fix $p \in [1, \infty)$, error probability $\alpha \in (0, 1)$, and suppose null hypothesis $H_0 :\ \mu = \nu$ and Assumption 1 holds. With probability at least $1 - \alpha$, it holds that*

$$\mathcal{KMS}_p(\widehat{\mu}_n, \widehat{\nu}_n) \le \Delta(n, \alpha).$$

The dimension free upper bound $\Delta(n,\alpha)=O(n^{-1(/2p)})$ is optimal in the worst case. Indeed, in the one-dimension case $\mathcal{B}=[0,1]$ and $K(x,y)=xy$, the kernel max-sliced Wasserstein distance $\mathcal{KMS}_p$ coincides with the classical Wasserstein distance W_p. In this case, it is easy to see that if $\mu=(\delta_0+\delta_1)/2$ is supported on the two points 0 and 1, then the expectation of $\mathcal{KMS}(\widehat{\mu}_n,\widehat{\nu}_n)$ is of order $n^{-1/(2p)}$ [21].

Suppose we design a two-sample test $\mathcal{T}_{\mathrm{KMS}}$ such that H_0 is rejected if $\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\geq\Delta(n,\alpha)$. By Theorem 1, we have the following performance guarantees of $\mathcal{T}_{\mathrm{KMS}}$.

COROLLARY 1 **(Testing Power of $\mathcal{T}_{\mathrm{KMS}}$).** *Fix a level $\alpha\in(0,1/2)$, $p\in[1,\infty)$, and suppose Assumption 1 holds. Then the following result holds:*

(I) *(Risk): The type-I risk of $\mathcal{T}_{\mathrm{KMS}}$ is at most α;*
(II) *(Power): Under $H_1:\ \mu\neq\nu$, suppose the sample size n is sufficiently large such that $\varrho_n:=\mathcal{KMS}_p(\mu,\nu)-\Delta(n,\alpha)>0$, the power of $\mathcal{T}_{\mathrm{KMS}}$ is at least $1-c\cdot n^{-1/(2p)}$, where c is a constant depending on A,C,p,ϱ_n.*

REMARK 3 (COMPARISION WITH MAXIMUM MEAN DISCREPANCY (MMD)). MMD has been a popular kernel-based tool to quantify the discrepancy between two probability measures (see, e.g., [26, 22, 35, 60, 61, 6, 48, 44, 63, 67]), which, for any two probability distributions μ and ν, is defined as

$$\mathrm{MMD}(\mu,\nu)=\max_{\substack{f\in\mathcal{H},\\ \|f\|_{\mathcal{H}}\leq 1}}\mathbb{E}_\mu[f]-\mathbb{E}_\nu[f]=\max_{\substack{f\in\mathcal{H},\\ \|f\|_{\mathcal{H}}\leq 1}}\overline{(u_f)_\#\big(\Phi_\#\mu\big)}-\overline{(u_f)_\#\big(\Phi_\#\nu\big)}, \tag{6}$$

where $\overline{\xi}$ denotes the mean of a given probability measure ξ. The empirical (biased) MMD estimator also exhibits dimension-free finite-sample guarantee as in Theorem 1: it decays in the order of $\mathcal{O}(n^{-1/2})$, where n is the number of samples. However, the KMS Wasserstein distance is more flexible as it replaces the mean difference objective in (6) by the Wasserstein distance, which naturally incorporates the geometry of the sample space and is suitable for hedging against adversarial data perturbations [23].

4. Computation of 2-KMS Wasserstein distance

Let $\widehat{\mu}_n$ and $\widehat{\nu}_n$ be two empirical distributions supported on n points, i.e., $\widehat{\mu}_n=\frac{1}{n}\sum_i\delta_{x_i},\widehat{\nu}_n=\frac{1}{n}\sum_j\delta_{y_j}$, where $\{x_i\}_i,\{y_j\}_j$ are data points in $\mathbb{R}^d$. This section focuses on the computation of 2-KMS Wasserstein distance between these two distributions. According to Definition 3, it holds that

$$\mathcal{KMS}_2(\widehat{\mu}_n,\widehat{\nu}_n)=\left(\max_{f\in\mathcal{H},\ \|f\|_{\mathcal{H}}^2\leq 1}\left\{\min_{\pi\in\Gamma_n}\sum_{i,j\in[n]}\pi_{i,j}|f(x_i)-f(y_j)|^2\right\}\right)^{1/2}, \tag{KMS}$$

where Γ_n is defined in (5).

Although the outer maximization problem is a *functional optimization* that contains uncountably many parameters, one can apply the existing represented theorem (see below) to reformulate Problem (KMS) as a finite-dimensional optimization.

THEOREM 2 **(Theorem 1 in [70]).** *There exists an optimal solution to (KMS), denoted as $\widehat{f}$, such that for any z,*

$$\widehat{f}(z)=\sum_{i=1}^n a_{x,i}K(z,x_i)-\sum_{i=1}^n a_{y,i}K(z,y_i), \tag{7}$$

where $a_x=(a_{x,i})_{i\in[n]},a_y=(a_{y,i})_{i\in[n]}$ are coefficients to be determined.

Define gram matrix $K(x^n, x^n) = (K(x_i, x_j))_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ and other gram matrices $K(x^n, y^n)$, $K(y^n, x^n), K(y^n, y^n)$ likewise, then define the concatenation of gram matrics

$$G = [K(x^n, x^n), -K(x^n, y^n); -K(y^n, x^n), K(y^n, y^n)] \in \mathbb{R}^{2n\times 2n}. \tag{8}$$

Assume G is positive definite such that it admits the Cholesky decomposition $G^{-1} = UU^{\mathrm{T}}$. By substituting the expression (7) into (KMS) and direct calculation (see Appendix EC.4), we obtain the following exact reformulation of (KMS):

$$\max_{\omega\in\mathbb{R}^{2n}:\ \|\omega\|_2=1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 \right\}. \tag{9}$$

Here, we omit taking the square root of the optimal value of the max-min optimization problem for simplicity of presentation and the vector

$$M_{i,j} = U^{\mathrm{T}} M'_{i,j}, \text{ where } M'_{i,j} = [(K(x_i, x_l) - K(y_j, x_l))_{l\in[n]}; (K(y_j, y_l) - K(x_i, y_l))_{l\in[n]}] \in \mathbb{R}^{2n}.$$

Since Problem (9) is a non-convex program, it is natural to question its computational hardness. The following theorem gives an affirmative answer, whose proof is provided in Appendix EC.5.

THEOREM 3 **(NP-hardness of computing 2-KMS Wasserstein)**. *Problem (9) is NP-hard for the worst-case instances of* $\{M_{i,j}\}_{i,j}$.

The proof idea of Theorem 3 is to construct an instance of $\{M_{i,j}\}_{i,j}$ that depends on a generic collection of n vectors $\{A_i\}_i$ such that solving (9) is at least as difficult as solving the optimization $\max_{\omega:\ \|\omega\|_2=1} \min_{i\in[n]} \omega^{\mathrm{T}} A_i A_i^{\mathrm{T}} \omega$, which is the fair-PCA problem [59] with rank-1 data matrices (or fair beamforming problem [62]) and has been proved to be NP-hard [62]. Interestingly, the computational hardness of MS Wasserstein distance arises both from the high data dimension d and large sample size n, whereas that of KMS Wasserstein distance arises from the large sample size n only.

To tackle the computational challenge of solving (9), in the subsequent subsections, we present an SDR formula and propose an efficient first-order algorithm to solve it. Next, we analyze the computational complexity of our proposed algorithm and the theoretical guarantees on SDR.

4.1. Semidefinite relaxation with efficient algorithms

We observe the simple transformation of the objective in (9):

$$\sum_{i,j} \pi_{i,j} (M_{i,j}^{\mathrm{T}}\omega)^2 = \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \omega\omega^{\mathrm{T}} \rangle.$$

Inspired by this relation, we use the change of variable approach to optimize the rank-1 matrix $S = \omega\omega^{\mathrm{T}}$, i.e., it suffices to consider the following equivalent formulation of (9):

$$\max_{S\in\mathbb{S}_+^{2n}, \mathrm{Trace}(S)=1, \mathrm{rank}(S)=1} \left\{ F(S) = \min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle \right\}. \tag{10}$$

An efficient SDR is to drop the rank-1 constraint to consider the semidefinite program (SDP):

$$\max_{S\in\mathcal{S}_{2n}} F(S), \quad \text{where} \quad \mathcal{S}_{2n} = \left\{ S \in \mathbb{S}_+^{2n} :\ \mathrm{Trace}(S) = 1 \right\}. \tag{SDR}$$

REMARK 4 (CONNECTION WITH [74]). We highlight that Xie and Xie [74] considered the same SDR heuristic to compute the max-sliced 1-Wasserstein distance. However, the authors therein apply the interior point method to solve a large-scale SDP, which has expansive complexity $\mathcal{O}(n^{6.5}\text{polylog}(\frac{1}{\delta}))$ (up to δ-accuracy) [5]. In the following, we present a first-order method that exhibits much smaller complexity $\widetilde{\mathcal{O}}\left(n^2\delta^{-3}\right)$ in terms of the problem size n (see Theorem 4). Besides, theoretical guarantees on the solution from SDR have not been explored in [74], and we are the first literature to provide these results.

The constraint set $\mathcal{S}_{2n}$ is called the *Spectrahedron* and admits closed-form Bregman projection. Inspired by this, we propose an inexact mirror ascent algorithm to solve (SDR). Its high-level idea is to iteratively construct an inexact gradient estimator and next perform the mirror ascent on iteration points. By properly balancing the trade-off between the bias and cost of querying gradient oracles, this type of algorithm could guarantee to find a near-optimal solution [28, 29, 27].

We first discuss how to construct supgradient estimators of F. By Danskin's theorem [7],

$$\partial F(S) = \text{Conv}\Big\{ \sum_{i,j} \pi^*_{i,j}(S) M_{i,j} M_{i,j}^{\mathrm{T}} : \quad \pi^*(S) \in \Pi(S) \Big\},$$

where $\Pi(S)$ denotes the set of optimal solutions to the following optimal transport (OT) problem:

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle. \tag{11}$$

Therefore, the main challenge of constructing a supgradient estimator is to compute an optimal solution $\pi^*(S) \in \Gamma(S)$. Since computing an exactly optimal solution is too expensive, we derive its near-optimal estimator, denoted as $\widehat{\pi}$, and practically use the following supgradient estimator:

$$v(S) = \sum_{i,j} \widehat{\pi}_{i,j} M_{i,j} M_{i,j}^{\mathrm{T}}. \tag{12}$$

In particular, we adopt the stochastic gradient-based algorithm with Katyusha momentum in [75] to compute a ϵ-optimal solution $\widehat{\pi}$ to (11). It achieves the state-of-the-art complexity $\widetilde{\mathcal{O}}\left(n^2\epsilon^{-1}\right)$. See the detailed algorithm in Appendix EC.6. Next, we describe the main algorithm for solving (SDR). Define the (negative) von Neumann entropy $h(S) = \sum_{i \in [2n]} \lambda_i(S) \log \lambda_i(S)$, where $\{\lambda_i(S)\}_i$ are the eignevalues of S, and define the von Neumann Bregman divergence

$$V(S_1, S_2) = h(S_1) - h(S_2) - \langle S_1 - S_2, \nabla h(S_2)^{\mathrm{T}} \rangle = \text{Trace}(S_1 \log S_1 - S_1 \log S_2).$$

Iteratively, we update S_{k+1} by performing mirror ascent with constant stepsize $\alpha > 0$:

$$S_{k+1} = \arg\max_{S \in \mathcal{S}_{2n}} \alpha \langle v(S_k), S \rangle + V(S, S_k),$$

which admits the following closed-form update:

$$\widetilde{S}_{k+1} = \exp\left(\log S_k + \alpha v(S_k)\right), \qquad S_{k+1} = \widetilde{S}_{k+1} / \text{Trace}(\widetilde{S}_{k+1}). \tag{13}$$

The general procedure for solving (SDR) is summarized in Algorithm 1.

Algorithm 1 Inexact Mirror Ascent for solving (SDR)

1: **Input:** Max iterations T, initial guess S_1, tolerance ϵ, constant stepsize α.
2: **for** $k = 1, \ldots, T-1$ **do**
3: Apply [75] to obtain a ϵ-optimal solution (denoted as $\widehat{\pi}$) to (11)
4: Construct inexact supgradient $v(S_k)$ according to (12)
5: Perform mirror ascent according to (13)
6: **end for**
7: **Return** $\widehat{S}_{1:T} = \frac{1}{T}\sum_{k=1}^{T} S_k$

4.2. Theoretical analysis

In this subsection, we provide the complexity of solving (SDR) and theoretical guarantees on the optimal solution of (SDR). It is worth mentioning that the constraint set $\mathcal{S}_{2n}$ is compact, and the objective in (SDR) is continuous, so an optimal solution, denoted as S^*, is guaranteed to exist and with finite optimal value. To analyze the complexity of Algorithm 1, we first derive the bias and computational cost of the supgradient estimator $v(S)$ in (12). Define the constant $C = \max_{i,j} \|M_{i,j}\|_2^2$.

LEMMA 1 **(Bias and Computational Cost)**. (I) ***(Bias)*** *$v(S)$ corresponds to the gradient of $\widehat{F}(S) = \sum_{i,j} \widehat{\pi}_{i,j}\langle M_{i,j}^{\mathrm{T}} M_{i,j}, S\rangle$, where $\widehat{\pi}$ is defined in (12) and $|F(S) - \widehat{F}(S)| \le \epsilon$;*
(II) ***(Cost)*** *The cost for computing (12) is $\mathcal{O}\left(C \cdot n^2\sqrt{\log n}\epsilon^{-1}\right)$, with $\mathcal{O}(\cdot)$ hiding some universal constant.*

Next, we analyze the error of the inexact mirror ascent framework in Algorithm 1.

LEMMA 2 **(Error Analysis of Algorithm 1)**. *When taking the stepsize $\alpha = \frac{\log(2n)}{C\sqrt{T}}$, the output $\widehat{S}_{1:T}$ from Algorithm 1 satisfies*

$$0 \le F(S^*) - F(\widehat{S}_{1:T}) \le 2\epsilon + 2C\sqrt{\frac{\log(2n)}{T}}.$$

Combining Lemmas 1 and 2, we obtain the complexity for solving (SDR).

THEOREM 4 **(Complexity Bound)**. *Fix the precision $\delta > 0$ and specify hyper-parameters*

$$T = \lceil \frac{16C^2\log(2n)}{\delta^2} \rceil, \qquad \epsilon = \frac{\delta}{4}, \qquad \alpha = \frac{\log(2n)}{C\sqrt{T}}.$$

Then, the total cost of Algorithm 1 for finding δ-optimal solution to (SDR) is

$$\mathcal{O}\left(T \cdot Cn^2\sqrt{\log n}\delta^{-1}\right) = \mathcal{O}\left(C^3 n^2 (\log n)^{3/2}\delta^{-3}\right).$$

Next, we analyze the quality of (SDR). Notably, the exact formulation (9) requires the optimal solution to be rank-1 while the more tractable relaxation (SDR) does not enforce such a constraint. Therefore, it is of interest to provide theoretical guarantees on the low-rank solution of (SDR), i.e., we aim to find the smallest integer $k \ge 1$ such that there exists an optimal solution to (SDR) that is at most rank-k. The integer k is called a rank bound on (SDR).

Theorem 5 below characterizes the value of k, and Fig. 1 illustrates the comparison between the theoretical rank k, the ideal rank 1 required by the exact formulation (9), and the trivial rank bound $2n$ (as the matrix S is of size $2n \times 2n$). From Theorem 5 and Fig. 1, we find our theoretical rank bound is remarkably smaller than the trivial rank $2n$ and relatively close to the ideal rank 1. The proof of Theorem 5 is provided in Appendix EC.10.

THEOREM 5 **(Rank Bound on (SDR))**. *There exists an optimal solution to (SDR) of rank at most $k \triangleq 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$. As a result,*

$$\textit{Optval}(9) = \max_{S \in \mathbb{S}_+^{2n}, \textit{Trace}(S)=1, \textit{rank}(S)=1} F(S) \le \textit{Optval}(SDR) \le \max_{S \in \mathbb{S}_+^{2n}, \textit{Trace}(S)=1, \textit{rank}(S)=k} F(S).$$

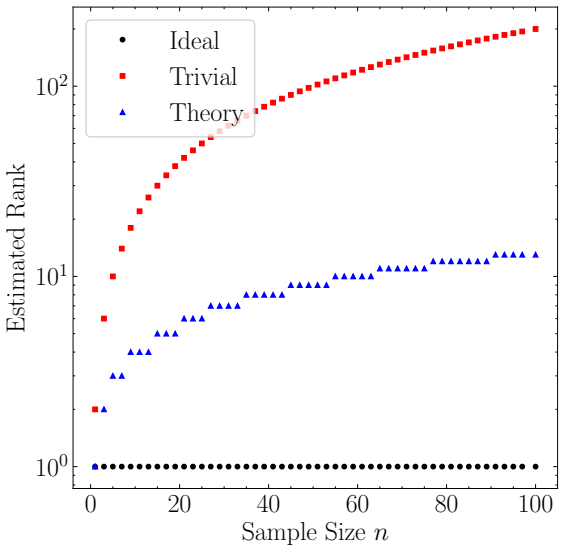

Figure 1 Comparision between the theoretical rank k, the ideal rank 1, and the trivial rank $2n$. The y-axis is in the logarithm scale.

Proof Sketch of Theorem 5. We first reformulate (SDR) by taking the dual of the inner OT problem:

$$\max_{\substack{S\in\mathcal{S}_{2n}\\ f,g\in\mathbb{R}^n}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(f_i+g_i): \quad f_i+g_j \le \langle M_{i,j}M_{i,j}^{\mathrm{T}}, S\rangle, \quad \forall i,j\in[n] \right\}. \tag{14}$$

By Birkhoff's theorem [8] and complementary slackness condition of OT, one can show that there exists an optimal solution of (f,g) such that at most n constraints of (14) are binding, and with such an optimal choice, one can adopt the convex geometry analysis from [39, 40] to derive the desired rank bound for any feasible extreme point of variable S. Since the set of optimal solutions of (SDR) must have a non-empty intersection with the set of feasible extreme points, the desired result holds. □

However, we highlight that Algorithm 1 only finds a near-optimal solution $\widehat{S}_{1:T}$ of (SDR), which is not guaranteed to satisfy the rank bound in Theorem 5. To fill the gap, we develop a rank-reduction algorithm that further converts $\widehat{S}_{1:T}$ to the solution that simultaneously maintains the desired rank bound and optimality gap. First, we fix $S\equiv\widehat{S}_{1:T}$ in (14) and find the optimal (f,g) with n binding constraints only, using the Hungarian algorithm [38]. Next, for fixed (f,g), we develop a greedy rank reduction algorithm inspired by [40, Algorithm 2], which iteratively reduces the rank of variable S until it reaches the desired rank bound. We provide the detailed algorithm description in Appendix EC.10 and complexity analysis below.

THEOREM 6. *There exists a rank-reduction algorithm (see Algorithm 4 in Appendix EC.10) such that* (I) *for a δ-optimal solution to (SDR), it outputs another δ-optimal solution with rank at most k;* (II) *its worst-case complexity is $\mathcal{O}(n^5)$.*

5. Numerical Study

This section presents experiment results for KMS 2-Wasserstein using SDR relaxation with first-order algorithm and rank reduction (denoted as `SDR-Efficient`). Baseline approaches include the block coordinate descent (`BCD`) algorithm [70], which finds stationary points of KMS 2-Wasserstein, and using off-the-shelf solver `cvxpy` [19] for solving SDR relaxation (denoted as `SDR-Naive`). All experiments were conducted on a MacBook Pro with an Intel Core i9 2.4GHz and 16GB memory, based on four datasets: `blob` (a d-dimensional Gaussian mixture synthetic dataset) [44], `Iris` [20], `mnist` [15],

and `credit` [76]. Unless otherwise stated, error bars are reproduced using 10 independent trials. Throughout the experiments, we specify the kernel as Gaussian, with bandwidth being the median of pairwise distances between data points. Other details and numerical studies can be found in Appendices EC.11 and EC.12.

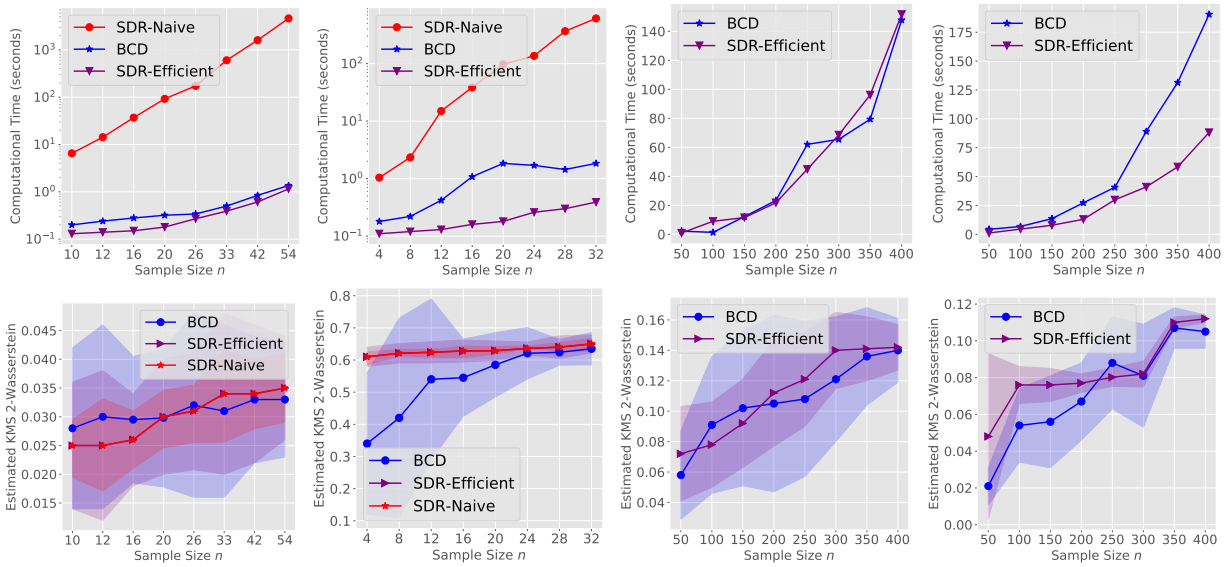

Figure 2 Comparison of `SDR-Efficiennt` with baselines `SDR-Naive` and `BCD` in terms of time and solution quality. Columns from left to right correspond to datasets `blob` (2-dimensional), `Iris`, `mnist`, and `credit`.

Computational Time and Solution Quality. We first compare our approach with baseline methods in terms of running time and solution quality. For a given nonlinear projector, its quality is estimated by projecting the testing data points from two groups and then computing their 2-Wasserstein distance. The experimental results are presented in Fig. 2, from which we can see that even for small sample size datasets, `SDR-Naive` takes considerably longer time than `SDR-Efficient` and `BCD`. So we omit the experiment of `SDR-Naive` for `mnist` and `credit` datasets. From the bottom plots of Fig. 2, we find the performance of two SDR solvers outperform `BCD`, as indicated by their larger means and smaller variations. One possible explanation is that `BCD` is designed to find a local optimum solution for the original non-convex problem, making it highly sensitive to the initial guess and potentially less effective in achieving optimal performance.

Testing Power of KMS Versus MS. Next, we compare the performance between KMS and MS Wasserstein distances for two-sample testing. Fig. 3 illustrates a toy example where μ and ν are generated from blob dataset and presents the contour plots of optimal projection functions estimated by computing these two distances. The plots show that KMS operates by identifying a central point and projecting each data point based on its distance from this central point. Subsequently, a two-sample test is conducted utilizing the Wasserstein distance between the projected data points as a statistic. In contrast, the MS Wasserstein distance appears to be less flexible, as depicted in the contour plot, where the projection function operates by linearly separating the sample space.

We also examine the testing power of $\mathcal{T}_{\text{KMS}}$ using `blob` dataset with dimension $d = 20$ in Fig. 4 for different choices of sample size $n \in \{20, 40, 80, 160, 180, 200\}$. The type-I error is controlled within $\alpha = 0.05$, and the error bar is generated using 20 independent trials. Compared with baseline approaches (`MMD-O` [44], `ME` [34], `MS` [69]), we find the KMS Wasserstein-based two-sample test distinguishes the differences between two Gaussian mixtures very well. Unlike Theorem 1 that determines the threshold $\Delta(n, \alpha)$ to reject H_0 based on a theoretical error bound, we use the bootstrap strategy to estimate such a threshold.

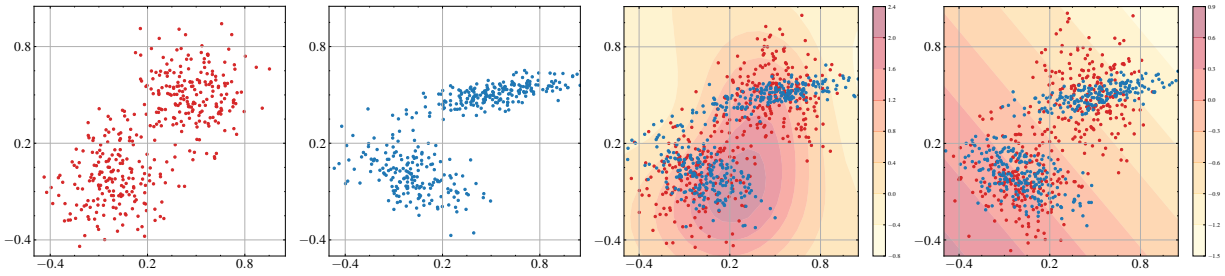

Figure 3 Visualization for 2-dimensional `blob` dataset. Figures from left to right correspond (a) samples from μ, (b) samples from ν under H_1, (c) optimal projector from KMS Wasserstein distance, and (d) optimal nonlinear projector from MS Wasserstein distance.

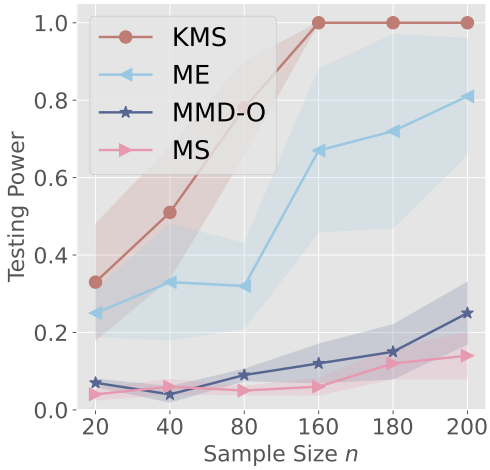

Figure 4 Testing power of KMS and other baseline approaches for `blob` dataset.

6. Concluding Remarks

In this paper, we presented statistical and computational guarantees of KMS Wasserstein distance. Our finite-sample guarantees demonstrate that the empirical KMS p-Wasserstein distance decays in the order of $n^{-1/(2p)}$ with n samples. Our findings are based on modest technical assumptions and do not face the curse of dimensionality. Regarding algorithms, we prove that computing KMS 2-Wasserstein distance between discrete measures is NP-hard. Subsequently, we introduce an effective semidefinite programming relaxation (SDR) and propose a first-order method utilizing biased gradient oracles to find its solution. Furthermore, we show that the SDR includes a solution of low rank and propose a greedy rank reduction algorithm that yields the desired low-rank solution. Finally, our numerical study validates our theoretical results and highlights the exceptional performance of the KMS Wasserstein distance.

References

[1] Altschuler J, Weed J, Rigollet P (2017) Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in Neural Information Processing Systems*, 1961–1971.

[2] Balagopalan A, Novikova J, Mcdermott MB, Nestor B, Naumann T, Ghassemi M (2020) Cross-language aphasia detection using optimal transport domain adaptation. *Machine Learning for Health Workshop*, 202–219 (PMLR).

[3] Bartl D, Mendelson S (2022) Structure preservation via the wasserstein distance. *arXiv preprint arXiv:2209.07058* .

[4] Barvinok AI (1995) Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry* 13:189–202.

[5] Ben-Tal A, Nemirovski A (2021) Lectures on modern convex optimization 2020. *SIAM, Philadelphia* .

[6] Berlinet A, Thomas-Agnan C (2011) *Reproducing kernel Hilbert spaces in probability and statistics* (Springer Science & Business Media).

[7] Bertsekas DP (1997) Nonlinear programming. *Journal of the Operational Research Society* 48(3):334–334.

[8] Birkhoff G (1946) Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A* 5:147–154.

[9] Boedihardjo MT (2024) Sharp bounds for max-sliced wasserstein distances. *arXiv preprint arXiv:2403.00666* .

[10] Bonneel N, Rabin J, Peyré G, Pfister H (2015) Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51:22–45.

[11] Carriere M, Cuturi M, Oudot S (2017) Sliced wasserstein kernel for persistence diagrams. *International conference on machine learning*, 664–673 (PMLR).

[12] Courty N, Flamary R, Habrard A, Rakotomamonjy A (2017) Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems* 30.

[13] Courty N, Flamary R, Tuia D (2014) Domain adaptation with regularized optimal transport. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, 274–289 (Springer).

[14] Courty N, Flamary R, Tuia D, Rakotomamonjy A (2016) Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 39(9):1853–1865.

[15] Deng L (2012) The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* 29(6):141–142.

[16] Deshpande I, Hu YT, Sun R, Pyrros A, Siddiqui N, Koyejo S, Zhao Z, Forsyth D, Schwing AG (2019) Max-sliced wasserstein distance and its use for gans. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10648–10656.

[17] Deshpande I, Zhang Z, Schwing AG (2018) Generative modeling using the sliced wasserstein distance. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3483–3491.

[18] Deza MM, Laurent M, Weismantel R (1997) *Geometry of cuts and metrics*, volume 2 (Springer).

[19] Diamond S, Boyd S (2016) Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17(83):1–5.

[20] Fisher RA (1988) Iris. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C56C76.

[21] Fournier N, Guillin A (2015) On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields* 162(3):707–738.

[22] Fukumizu K, Gretton A, Lanckriet G, Schölkopf B, Sriperumbudur BK (2009) Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in neural information processing systems* 22.

[23] Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.

[24] Gao R, Xie L, Xie Y, Xu H (2018) Robust hypothesis testing using wasserstein uncertainty sets. *Advances in Neural Information Processing Systems* 31.

[25] Genevay A, Peyré G, Cuturi M (2018) Learning generative models with sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics*, 1608–1617 (PMLR).

[26] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. *J. Mach. Learn. Res.* 13(null):723–773, ISSN 1532-4435.

[27] Hu Y, Chen X, He N (2021) On the bias-variance-cost tradeoff of stochastic optimization. *Advances in Neural Information Processing Systems* 34:22119–22131.

[28] Hu Y, Wang J, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems* 36.

[29] Hu Y, Zhang S, Chen X, He N (2020) Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems* 33:2759–2770.

[30] Huang M, Ma S, Lai L (2021) Projection robust wasserstein barycenters. *International Conference on Machine Learning*, 4456–4465 (PMLR).

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

[32] Huang M, Ma S, Lai L (2021) A riemannian block coordinate descent method for computing the projection robust wasserstein distance. *Proceedings of the 38th International Conference on Machine Learning*, 4446–4455.

[33] Jiang B, Liu YF (2024) A riemannian exponential augmented lagrangian method for computing the projection robust wasserstein distance. *Advances in Neural Information Processing Systems* 36.

[34] Jitkrittum W, Szabó Z, Chwialkowski KP, Gretton A (2016) Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems* 29.

[35] Kirchler M, Khorasani S, Kloft M, Lippert C (2020) Two-sample testing using deep learning. *International Conference on Artificial Intelligence and Statistics*, 1387–1398 (PMLR).

[36] Kolouri S, Nadjahi K, Simsekli U, Badeau R, Rohde G (2019) Generalized sliced wasserstein distances. *Advances in neural information processing systems* 32.

[37] Kolouri S, Pope PE, Martin CE, Rohde GK (2018) Sliced wasserstein auto-encoders. *International Conference on Learning Representations*.

[38] Kuhn HW (1955) The hungarian method for the assignment problem. *Naval research logistics quarterly* 2(1-2):83–97.

[39] Li Y, Xie W (2022) On the exactness of dantzig-wolfe relaxation for rank constrained optimization problems. *arXiv preprint arXiv:2210.16191* .

[40] Li Y, Xie W (2024) On the partial convexification for low-rank spectral optimization: Rank bounds and algorithms. *arXiv preprint arXiv:2305.07638, Forthcoming at Integer Programming and Combinatorial Optimization* .

[41] Lin T, Fan C, Ho N, Cuturi M, Jordan M (2020) Projection robust wasserstein distance and riemannian optimization. *Advances in neural information processing systems* 33:9383–9397.

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

[43] Lin T, Zheng Z, Chen E, Cuturi M, Jordan MI (2021) On projection robust optimal transport: Sample complexity and model misspecification. *International Conference on Artificial Intelligence and Statistics*, 262–270 (PMLR).

[44] Liu F, Xu W, Lu J, Zhang G, Gretton A, Sutherland DJ (2020) Learning deep kernels for non-parametric two-sample tests. *International Conference on Machine Learning*, 6316–6326.

[45] Luise G, Rudi A, Pontil M, Ciliberto C (2018) Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems* 31.

[46] Mairal J, Vert JP (2018) Machine learning with kernels. mines paristech, paris, france. URL https://members.cbio.mines-paristech.fr/~jvert/svn/kernelcourse/slides/master2017/master2017.pdf.

[47] Mika S, Schölkopf B, Smola A, Müller KR, Scholz M, Rätsch G (1998) Kernel pca and de-noising in feature spaces. *Advances in neural information processing systems* 11.

[48] Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B, et al. (2017) Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* 10(1-2):1–141.

[49] Nadjahi K, Durmus A, Simsekli U, Badeau R (2019) Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems* 32.

[50] Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* 19(4):1574–1609.

[51] Nguyen K, Ho N (2023) Energy-based sliced wasserstein distance. *Advances in Neural Information Processing Systems* 36.

[52] Nietert S, Goldfeld Z, Sadhu R, Kato K (2022) Statistical, robustness, and computational guarantees for sliced wasserstein distances. *Advances in Neural Information Processing Systems* 35:28179–28193.

[53] Pataki G (1998) On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research* 23(2):339–358.

[54] Patrini G, Van den Berg R, Forre P, Carioni M, Bhargav S, Welling M, Genewein T, Nielsen F (2020) Sinkhorn autoencoders. *Uncertainty in Artificial Intelligence*, 733–743 (PMLR).

[55] Paty FP, Cuturi M (2019) Subspace robust wasserstein distances. *International conference on machine learning* 5072–5081.

[56] Petzka H, Fischer A, Lukovnikov D (2018) On the regularization of wasserstein gans. *International Conference on Learning Representations*.

[57] Peyre G, Cuturi M (2019) Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning* 11(5-6):355–607.

[58] Ramdas A, García Trillos N, Cuturi M (2017) On wasserstein two-sample testing and related families of nonparametric tests. *Entropy* 19(2):47.

[59] Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S (2018) The price of fair pca: One extra dimension. *Advances in neural information processing systems* 31.

[60] Schrab A, Kim I, Albert M, Laurent B, Guedj B, Gretton A (2021) Mmd aggregated two-sample test. *arXiv preprint arXiv:2110.15073* .

[61] Schrab A, Kim I, Guedj B, Gretton A (2022) Efficient aggregated kernel tests using incomplete u-statistics. *Advances in Neural Information Processing Systems* 35:18793–18807.

[62] Sidiropoulos ND, Davidson TN, Luo ZQ (2006) Transmit beamforming for physical-layer multicasting. *IEEE transactions on signal processing* 54(6):2239–2251.

[63] Sutherland DJ, Tung HY, Strathmann H, De S, Ramdas A, Smola A, Gretton A (2016) Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488* .

[64] Tantipongpipat U, Samadi S, Singh M, Morgenstern JH, Vempala S (2019) Multi-criteria dimensionality reduction with applications to fairness. *Advances in neural information processing systems* 32.

[65] Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).

[66] Wang J, Chen M, Zhao T, Liao W, Xie Y (2023) A manifold two-sample test study: integral probability metric with neural networks. *Information and Inference: A Journal of the IMA* 12(3):1867–1897.

[67] Wang J, Dey SS, Xie Y (2023) Variable selection for kernel two-sample tests. *arXiv preprint arXiv:2302.07415* .

[68] Wang J, Gao R, Xie Y (2021) Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926* .

[69] Wang J, Gao R, Xie Y (2021) Two-sample test using projected wasserstein distance. *2021 IEEE International Symposium on Information Theory (ISIT)*, 3320–3325 (IEEE).

[70] Wang J, Gao R, Xie Y (2022) Two-sample test with kernel projected wasserstein distance. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151.

[71] Wang J, Gao R, Xie Y (2024) Non-convex robust hypothesis testing using sinkhorn uncertainty sets. *arXiv preprint arXiv:2403.14822* .

[72] Wang J, Moore R, Xie Y, Kamaleswaran R (2022) Improving sepsis prediction model generalization with optimal transport. *Machine Learning for Health*, 474–488 (PMLR).

[73] Wang J, Xie Y (2022) A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets. *2022 IEEE International Symposium on Information Theory (ISIT)*, 3315–3320 (IEEE).

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

[75] Xie Y, Luo Y, Huo X (2022) An accelerated stochastic algorithm for solving the optimal transport problem. *arXiv preprint arXiv:2203.00813* .

[76] Yeh IC (2016) Default of Credit Card Clients. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C55S3H.

Supplementary for *"Statistical and Computational Guarantees of Kernel Max-Sliced Wasserstein Distances"*

EC.1. Limitations

We list several limitations of this work below.

(I) Our theoretical results assume that the sample sizes from sample sets x^n and y^m satisfy $m = n$ for the simplicity of presentation. If considering the unequal sample size $m \neq n$, the error bound $\Delta(n, \alpha)$ in Theorem 1 is replaced by

$$2A\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p} \cdot \left[n^{-1/(2p)} + m^{-1/(2p)}\right];$$

the complexity bound in Theorem 4 replaces n with $\max\{n, m\}$; the rank bound in Theorem 5 replaces $k = 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$ with

$$1 + \left\lfloor \sqrt{2(n + m - 1) + \frac{9}{4}} - \frac{3}{2} \right\rfloor,$$

since for unbalanced OT, at most $n + m - 1$ constraints are binding [57, Proposition 3.4].

(II) Although our statistical guarantees focus on KMS p-Wasserstein distance with general $p \in [1, \infty)$, our computational guarantees focus on $p = 2$ only. It is of research interest to develop similar computational guarantees for other choices of p, i.e., $p = 1$.

(III) It is of research interest to develop performance guarantees on SDR in terms of optimal value instead of solution rank. Also, it is desirable to develop exact algorithms or approximation algorithms with better approximation quality for computing KMS p-Wasserstein distance.

(IV) In this work we only consider classical Gaussian kernel for numerical study. It is desirable to consider deep neural network-based kernel to further boost the power of KMS Wasserstein-based testing.

EC.2. Broader Societal Impact

This is a theoretical work on statistical and computational guarantees on KMS Wasserstein distance. One of its societal impacts is its application to non-parametric two-sample testing. In practice, researchers can deploy two-sample testing to evaluate the effectiveness of medical treatments, discover economic disparities, detect anomaly observations, and more. We do not foresee any negative societal impact of this work.

EC.3. Proof of Theorem 1 and Corollary 1

The proof in this part relies on the following technical results.

THEOREM EC.1. ***(Finite-Sample Guarantee on MS 1-Wasserstein Distance on Hilbert Space, Adopted from [9, Corollary 2.8])*** *Let $\delta \in (0,1]$, and μ be a probability measure on a separable Hilbert space $\mathcal{H}$ with $\int_{\mathcal{H}} \|x\| \, \mathrm{d}\mu(x) < \infty$. Let $X_1, \ldots, X_n$ be i.i.d. random elements of $\mathcal{H}$ sampled according to μ, and $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, then it holds that*

$$\mathbb{E}\mathcal{MS}_1\left(\mu, \widehat{\mu}_n\right) \leq C \cdot \left(\int_{\mathcal{H}} \|x\|^{2+2\delta} \, \mathrm{d}\mu(x)\right)^{1/(2+2\delta)} \cdot (\delta n)^{-1/2},$$

where $C \geq 1$ is a universal constant.

THEOREM EC.2 **(Functional Hoefffding Theorem [65, Theorem 3.26])**. *Let $\mathcal{F}$ be a class of functions, each of the form $h: \mathcal{B} \to \mathbb{R}$, and $X_1, \ldots, X_n$ be samples i.i.d. drawn from μ on $\mathcal{B}$. For $i \in [n]$, assume there are real numbers $a_{i,h} \leq b_{i,h}$ such that*

$$h(x) \in [a_{i,h}, b_{i,h}]$$

for any $x \in \mathcal{B}, h \in \mathcal{F} \cup \{-\mathcal{F}\}$. Define

$$L^2 = \sup_{h \in \mathcal{F} \cup \{-\mathcal{F}\}} \frac{1}{n}\sum_{i=1}^n (b_{i,h} - a_{i,h})^2.$$

For all $\delta \geq 0$, it holds that

$$\mathbb{P}\left\{\sup_{h\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n h(X_i)\right| \geq \mathbb{E}\left[\sup_{h\in\mathcal{F}} \left|\frac{1}{n}\sum_{1}^n h(X_i)\right|\right] + \delta\right\} \leq \exp\left(-\frac{n\delta^2}{4L^2}\right).$$

We first show the one-sample guarantees for KMS p-Wasserstein distance.

PROPOSITION EC.1. *Fix $p \in [1, \infty)$, error probability $\alpha \in (0,1)$, and suppose Assumption 1 holds. Let $C \geq 1$ be a universal constant. Then, we have the following results:*

(I) $\mathbb{E}\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \leq A(2C^{1/p}) \cdot n^{-1/(2p)}$

(II) *With probability at least $1-\alpha$, it holds that*

$$\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \leq 2^{1-1/p} A \left(C + 4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p} \cdot n^{-1/(2p)}.$$

Proof of Proposition EC.1. Recall from (3) that

$$\mathcal{KMS}_p(\mu, \nu) = \mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\nu\Big).$$

Therefore, it suffices to derive one-sample guarantees for $\mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big)$.

(I) Observe that under Assumption 1, we have

$$A^2 \geq K(x,x) = \langle K_x, K_x\rangle = \|K_x\|_{\mathcal{H}}^2,$$

and therefore $\|\Phi(x)\|_{\mathcal{H}} = \|K_x\|_{\mathcal{H}} \leq A, \forall x \in \mathcal{B}$. In other words, for every probability measure μ on $\mathcal{B}$, the probability measure $\Phi_{\#}\mu$ is supported on the ball in $\mathcal{H}$ centered at the origin with radius A. By Theorem EC.1 with $\delta = 1$, we obtain

$$\mathbb{E}\mathcal{KMS}_1(\mu, \widehat{\mu}_n) = \mathbb{E}\mathcal{MS}_1\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \leq \frac{AC}{\sqrt{n}}.$$

Since $\Phi_{\#}\mu$ and $\Phi_{\#}\widehat{\mu}_n$ are supported on the ball of $\mathcal{H}$ centered at the origin with radius A, it holds that

$$\mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \le \Big[\mathcal{MS}_1\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \cdot (2A)^{p-1}\Big]^{1/p}.$$

In other words,

$$\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \le \Big[\mathcal{KMS}_1(\mu, \widehat{\mu}_n) \cdot (2A)^{p-1}\Big]^{1/p}. \tag{EC.1}$$

It follows that

$$\begin{aligned}
\mathbb{E}\mathcal{KMS}_p(\mu, \widehat{\mu}_n) &= \mathbb{E}\mathcal{MS}_p\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \\
&\le \mathbb{E}\Big[\mathcal{MS}_1\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \cdot (2A)^{p-1}\Big]^{1/p} \\
&\le \Big\{\mathbb{E}\Big[\mathcal{MS}_1\Big(\Phi_{\#}\mu, \Phi_{\#}\widehat{\mu}_n\Big) \cdot (2A)^{p-1}\Big]\Big\}^{1/p} \\
&\le \left\{\frac{AC}{\sqrt{n}} \cdot (2A)^{p-1}\right\}^{1/p} = 2^{1-1/p} A C^{1/p} \cdot n^{-1/(2p)}.
\end{aligned}$$

(II) For the second part, we re-write $\mathcal{KMS}_1(\mu, \widehat{\mu}_n)$ with $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ using the kantorovich dual reformulation of OT:

$$\mathcal{KMS}_1(\mu, \widehat{\mu}_n) = \sup_{\substack{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\le 1,\\ g \text{ is 1-Lipschitz with } g(0)=0}} \left|\frac{1}{n}\sum_{i=1}^n \Big(g(f(x)) - \mathbb{E}_{x\sim\mu}[g(f(x))]\Big)\right|,$$

where the additional constraint $g(0) = 0$ does not impact the optimal value of the OT problem. In other words, one can represent

$$\mathcal{KMS}_1(\mu, \widehat{\mu}_n) = \sup_{h\in\mathfrak{H}} \Big|\frac{1}{n}\sum_{i=1}^n h(x_i)\Big|,$$

where the function class

$$\mathfrak{H} = \Big\{x \mapsto g(f(x)) - \mathbb{E}_{x\sim\mu}[g(f(x))] :\ g \text{ is 1-Lipschitz with } g(0) = 0, \quad f \in \mathcal{H}, \|f\|_{\mathcal{H}} \le 1\Big\}.$$

Consequently, for any x,

$$|g(f(x))| = |g(f(x)) - g(0)| \le |f(x)| = |\langle f, K_x\rangle_{\mathcal{H}}| \le \|f\|_{\mathcal{H}}\|K_x\|_{\mathcal{H}} \le A.$$

One can apply Theorem EC.2 with $\mathcal{F} \equiv \mathfrak{H}$, $a_{i,h} \equiv -A - \mathbb{E}_{x\sim\mu}[g(f(x))]$, $b_{i,h} \equiv A - \mathbb{E}_{x\sim\mu}[g(f(x))]$, where $h(x) = g(f(x)) - \mathbb{E}_{x\sim\mu}[g(f(x))]$, to obtain

$$\mathbb{P}\Big\{\mathcal{KMS}_1(\mu, \widehat{\mu}_n) \ge \mathbb{E}\left[\mathcal{KMS}_1(\mu, \widehat{\mu}_n)\right] + \delta\Big\} \le \exp\left(-\frac{n\delta^2}{4(2A)^2}\right) = \exp\left(-\frac{n\delta^2}{16A^2}\right).$$

Or equivalently, the following relation holds with probability at least $1-\alpha$:

$$\mathcal{KMS}_1(\mu, \widehat{\mu}_n) \le \mathbb{E}\left[\mathcal{KMS}_1(\mu, \widehat{\mu}_n)\right] + 4An^{-1/2}\sqrt{\log\frac{1}{\alpha}} \le An^{-1/2}\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right).$$

By the relation (EC.1), we find that with probability at least $1-\alpha$,

$$\begin{aligned}
&\mathcal{KMS}_p(\mu, \widehat{\mu}_n) \\
&\le \Big[An^{-1/2}\left(C + 4\sqrt{\log\frac{1}{\alpha}}\right) \cdot (2A)^{p-1}\Big]^{1/p} = 2^{1-1/p} A \left(C + 4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p} \cdot n^{-1/(2p)}.
\end{aligned}$$

□

We now complete the proof of Theorem 1. By the triangle inequality, with probability at least $1-2\alpha$, it holds that

$$\begin{aligned}\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n) &\le \mathcal{KMS}_p(\mu,\widehat{\mu}_n)+\mathcal{KMS}_p(\nu,\widehat{\nu}_n)\\ &\le 2\cdot 2^{1-1/p}A\left(C+4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p}\cdot n^{-1/(2p)}\\ &\le 4A\left(C+4\sqrt{\log\frac{1}{\alpha}}\right)^{1/p}\cdot n^{-1/(2p)}.\end{aligned}$$

Then, substituting α with $\alpha/2$ gives the desired result.

Proof of Corollary 1. It remains to show the type-II risk when proving this corollary. In particular,

$$\begin{aligned}\text{Type-II Risk} &= \mathbb{P}_{H_1}\left\{\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)<\Delta(n,\alpha)\right\}\\ &= \mathbb{P}_{H_1}\left\{\mathcal{KMS}_p(\mu,\nu)-\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\ge \mathcal{KMS}_p(\mu,\nu)-\Delta(n,\alpha)\right\}\\ &\le \mathbb{P}_{H_1}\left\{\left|\mathcal{KMS}_p(\mu,\nu)-\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\right|\ge \mathcal{KMS}_p(\mu,\nu)-\Delta(n,\alpha)\right\}\\ &\le \frac{\mathbb{E}\left|\mathcal{KMS}_p(\mu,\nu)-\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\right|}{\mathcal{KMS}_p(\mu,\nu)-\Delta(n,\alpha)},\end{aligned}$$

where the last relation is based on the Markov inequality and the assumption that $\mathcal{KMS}_p(\mu,\nu)-\Delta(n,\alpha)>0$. Based on the triangular inequality, we can see that

$$\mathbb{E}\left|\mathcal{KMS}_p(\mu,\nu)-\mathcal{KMS}_p(\widehat{\mu}_n,\widehat{\nu}_n)\right|\le \mathbb{E}[\mathcal{KMS}_p(\mu,\widehat{\mu}_n)]+\mathbb{E}[\mathcal{KMS}_p(\nu,\widehat{\nu}_n)]\le 2AC^{1/p}\cdot n^{-1/(2p)}.$$

Combining these two upper bounds, we obtain the desired result. □

EC.4. Reformulation for 2-KMS Wasserstein Distance in (KMS)

In this section, we derive the reformulation for computing 2-KMS Wasserstein distance:

$$\max_{f\in\mathcal{H},\ \|f\|_{\mathcal{H}}^2\le 1}\left\{\min_{\pi\in\Gamma_n}\sum_{i,j\in[n]}\pi_{i,j}|f(x_i)-f(y_j)|^2\right\}. \tag{EC.2}$$

Based on the expression of f in (7), we reformulate the problem above as

$$\max_{a_x,a_y\in\mathbb{R}^n}\left\{\min_{\pi\in\Gamma_n}\sum_{i,j\in[n]}\pi_{i,j}\left|\sum_{l\in[n]}a_{x,l}K(x_i,x_l)-\sum_{l\in[n]}a_{y,l}K(y_j,y_l)\right|^2\right\}, \tag{EC.3a}$$

subject to the constraint

$$\begin{aligned}&\left\|\sum_{i=1}^n a_{x,i}K(\cdot,x_i)-\sum_{i=1}^n a_{y,i}K(\cdot,y_i)\right\|_{\mathcal{H}}^2\\ =&\langle\sum_{i=1}^n a_{x,i}K(\cdot,x_i)-\sum_{i=1}^n a_{y,i}K(\cdot,y_i)\sum_{i=1}^n a_{x,i}K(\cdot,x_i)-\sum_{i=1}^n a_{y,i}K(\cdot,y_i)\rangle\\ =&\sum_{i,j\in[n]}a_{x,i}a_{x,j}\langle K(\cdot,x_i),K(\cdot,x_j)\rangle+\sum_{i,j\in[n]}a_{y,i}a_{y,j}\langle K(\cdot,y_i),K(\cdot,y_j)\rangle\\ &-2\sum_{i,j\in[n]}a_{x,i}a_{y,j}\langle K(\cdot,x_i),K(\cdot,y_j)\rangle\le 1.\end{aligned} \tag{EC.3b}$$

One can re-write (EC.3) in compact matrix form. If we define

$$\begin{aligned}s&=[a_x;a_y],\\ M'_{i,j}&=[(K(x_i,x_l)-K(y_i,x_l))_{l\in[n]};(K(y_j,y_l)-K(x_i,y_l))_{l\in[n]}],\\ G&=[K(x^n,x^n),-K(x^n,y^n);-K(y^n,x^n),K(y^n,y^n)]\in\mathbb{R}^{2n\times 2n},\end{aligned}$$

Problem (EC.3) can be reformualted as

$$\max_{s\in\mathbb{R}^{2n}}\left\{\min_{\pi\in\Gamma_n}\sum_{i,j\in[n]}\pi_{i,j}\left|s^{\mathrm{T}}M'_{i,j}\right|^2:\quad s^{\mathrm{T}}Gs\le 1\right\}. \tag{EC.4}$$

Take Cholesky decomposition $G^{-1}=UU^{\mathrm{T}}$ and use the change of variable approach to take $\omega=U^{-1}s$, Problem (EC.4) can be further reformulated as

$$\max_{s\in\mathbb{R}^{2n}}\left\{\min_{\pi\in\Gamma_n}\sum_{i,j\in[n]}\pi_{i,j}\left(\langle\omega,U^{\mathrm{T}}M'_{i,j}\rangle\right)^2:\quad \omega^{\mathrm{T}}\omega\le 1\right\}. \tag{EC.5}$$

After defining $M_{i,j}=U^{\mathrm{T}}M'_{i,j}$ and observing that the inequality constraint $\omega^{\mathrm{T}}\omega\le 1$ will become tight, we obtain the desired reformulation as in (9).

EC.5. Proof of Theorem 3

The general procedure of NP-hardness proof is illustrated in the following diagram: Problem (9) contains the **(Fair PCA with rank-1 data)** as a special case, whereas this special problem further contains **(Partition)** (which is known to be NP-complete) as a special case. After building these two reductions, we finish the proof of Theorem 3.

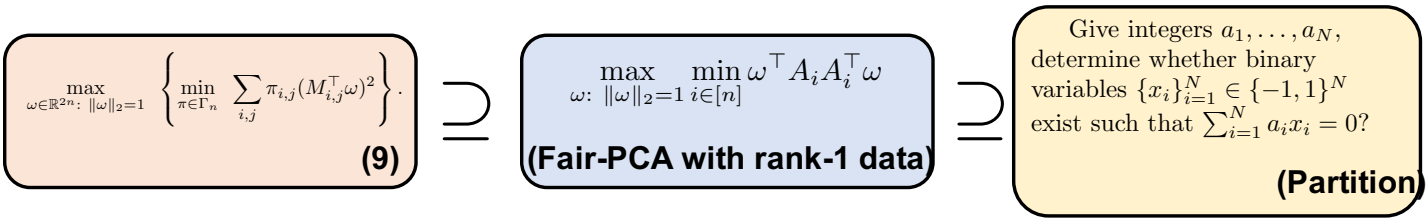

Figure EC.1 Proof outline of Theorem 3

Claim 1. Problem (9) contains Problem (**Fair PCA with rank-1 data**).

Proof of Claim 1. Given vectors $A_1,\ldots,A_n$, we specify

$$M_{1,:} \triangleq \{M_{1,1}, M_{1,2}, \ldots, M_{1,n}\} = \{A_1,\ldots,A_n\},$$

and $M_{i,:} \triangleq \{M_{i,1}, M_{i,2}, \ldots, M_{i,n}\}, i = 2,\ldots,n$ is specified by circularly shifting the elements in $M_{1,:}$ by $i-1$ positions. For instance, $M_{2,:} = \{A_n, A_1, \ldots, A_{n-1}\}$. For the inner OT problem in (9), it suffices to consider deterministic optimal transport π, i.e.,

$$\pi_{i,j} = \begin{cases} 1/n, & \text{if } j = \sigma(i), \\ 0, & \text{otherwise} \end{cases}$$

for some bijection mapping $\sigma:\ [n] \to [n]$. The cost matrix for the inner OT is actually a circulant matrix:

$$\left((M_{i,j}^{\mathrm{T}}\omega)^2\right)_{i,j} = \begin{pmatrix} (A_1\omega)^2 & (A_2\omega)^2 & \cdots & (A_n\omega)^2 \\ (A_n\omega)^2 & (A_1\omega)^2 & \cdots & (A_{n-1}\omega)^2 \\ \vdots & \vdots & \ddots & \vdots \\ (A_2\omega)^2 & (A_3\omega)^2 & \cdots & (A_1\omega)^2 \end{pmatrix}.$$

When considering the feasible circularly shifting bijection mapping (e.g., $\sigma(i) = (i+j) \mod n, \forall i \in [n]$ for $j = 0,1,\ldots,n-1$), we obtain the upper bound on the optimal value of the inner OT problem in (9):

$$\min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j}(M_{i,j}^{\mathrm{T}}\omega)^2 \le \min_{i\in[n]} (A_i^{\mathrm{T}}\omega)^2 = \min_{i\in[n]} \omega^{\mathrm{T}} A_i A_i^{\mathrm{T}} \omega.$$

On the other hand, for any bijection mapping σ, the objective of the inner OT problem in (9) can be written as a convex combination of $(A_1^{\mathrm{T}}\omega)^2,\ldots,(A_n^{\mathrm{T}}\omega)^2$, and thus,

$$\min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j}(M_{i,j}^{\mathrm{T}}\omega)^2 \ge \min_{\alpha\in\mathbb{R}_n^+, \sum_i \alpha_i = 1} \left\{ \sum_i \alpha_i (A_i^{\mathrm{T}}\omega)^2 \right\} \ge \min_{i\in[n]} (A_i^{\mathrm{T}}\omega)^2.$$

Since the upper and lower bounds match with each other, we obtain

$$\min_{\pi\in\Gamma_n} \sum_{i,j} \pi_{i,j}(M_{i,j}^{\mathrm{T}}\omega)^2 = \min_{i\in[n]} \omega^{\mathrm{T}} A_i A_i^{\mathrm{T}} \omega,$$

and consequently,

$$\max_{\omega:\ \|\omega\|_2=1}\left\{\min_{\pi\in\Gamma_n}\ \sum_{i,j}\pi_{i,j}(M_{i,j}^{\mathrm{T}}\omega)^2\right\}=\max_{\omega:\ \|\omega\|_2=1}\left\{\min_{i\in[n]}\ \omega^{\mathrm{T}}A_iA_i^{\mathrm{T}}\omega\right\},$$

which justifies Problem (9) contains Problem (**Fair PCA with rank-1 data**). □

Claim 2. Problem (**Fair PCA with rank-1 data**) contains Problem (**Partition**).

It is noteworthy that Claim 2 has previously been proved by [62]. For the sake of completeness, we provide the proof here.

Proof of Claim 2. Consider the norm minimization problem

$$P=\min_{\omega}\ \left\{\|\omega\|_2^2:\ \min_{i\in[n]}(\omega^{\mathrm{T}}A_i)^2\geq 1\right\}. \tag{EC.6}$$

and the scaled problem

$$\max_{\omega}\ \left\{\min_{i\in[n]}\ (\omega^{\mathrm{T}}A_i)^2:\ \|\omega\|_2^2=P\right\}. \tag{EC.7}$$

We can show that Problem (**Fair PCA with rank-1 data**) is equivalent to (EC.7), whereas (EC.7) is equivalent to (EC.6). Indeed,

- For the first argument, for any optimal solution from Problem (**Fair PCA with rank-1 data**), denoted as ω^*, one can do the scaling to consider $\tilde{\omega}^*=\sqrt{P}\omega$, which is also optimal to (EC.7), and vise versa.
- For the second argument, let ω_1,ω_2 be optimal solutions from (EC.6), (EC.7), respectively. Since P is the optimal value of (EC.6), one can check that ω_1 is a feasible solution to (EC.7). Since $\min_{i\in[n]}\ (\omega_1^{\mathrm{T}}A_i)^2\geq 1$, by the optimality of ω_2, it holds that $\min_{i\in[n]}\ (\omega_2^{\mathrm{T}}A_i)^2\geq 1$, i.e., ω_2 is a feasible solution to (EC.6). Since $\|\omega_2\|_2^2=P$, ω_2 is an optimal solution to (EC.6). Reversely, one can show ω_1 is an optimal solution to (EC.7): suppose on the contrary that there exists $\bar{\omega}_1$ such that $\|\bar{\omega}_1\|_2^2=P$ and $\min_{i\in[n]}\ (\bar{\omega}_1^{\mathrm{T}}A_i)^2>\min_{i\in[n]}\ (\omega_1^{\mathrm{T}}A_i)^2\geq 1$, then one can do a scaling of $\bar{\omega}_1$ such that $\min_{i\in[n]}\ (\bar{\omega}_1^{\mathrm{T}}A_i)^2=1$ whereas $\|\bar{\omega}_1\|_2^2>P$, which contradicts to the optimality of P. Combining both directions, we obtain the equivalence argument.

Thus, it suffices to show (EC.6) contains Problem (**Partition**). Define $a=(a_i)_{i\in[n]}, Q=I_n+aa^{\mathrm{T}}$, and assume Q admits Cholesky factorization $Q=S^{\mathrm{T}}S$. Then we create the vector $A_i=S^{-\mathrm{T}}e_i$, where e_i is the i-th unit vector of length n. Then, it holds that

$$\begin{aligned}
&\text{(EC.6)}\\
=&\min_{\omega}\ \left\{\|\omega\|_2^2:\ \min_{i\in[n]}((S^{-1}\omega)^{\mathrm{T}}e_i)^2\geq 1\right\}\\
=&\min_{\omega}\ \left\{\|Sx\|_2^2:\ \min_{i\in[n]}(x^{\mathrm{T}}e_i)^2\geq 1\right\}\\
=&\min_{\omega}\ \left\{x^{\mathrm{T}}Qx:\ x_i^2\geq 1\right\}\\
=&\min_{\omega}\ \left\{\sum_{i=1}^n x_i^2+\left(\sum_{i=1}^n a_ix_i\right)^2:\ x_i^2\geq 1\right\} \qquad (*)
\end{aligned}$$

where the second equality is by the change of variable $x=S^{-1}\omega$, the third equality is by the definitions of S and e_i, and the last equality is by the definition of Q. The solution to Problem (**Partition**) exists if and only if the optimal value to Problem (*) equals n. Thus, we finish the proof of Claim 2.

EC.6. Algorithm that Finds Near-optimal Solution to Optimal Transport

In this section, we present the algorithm that returns ϵ-optimal solution to the following optimal transport (OT) problem:

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} c_{i,j}, \tag{EC.8}$$

where $\{c_{i,j}\}_{i,j}$ is the given cost matrix. Define $\|C\|_\infty = \max_{i,j} c_{i,j}$. In other words, we find $\widehat{\pi} \in \Gamma_n$ such that

$$\texttt{optval(EC.8)} \le \sum_{i,j} \widehat{\pi}_{i,j} c_{i,j} \le \texttt{optval(EC.8)} + \epsilon.$$

Entropy-Regularized OT. The key to the designed algorithm is to consider the entropy regularized OT problem

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} c_{i,j} + \eta \sum_{i,j} \pi_{i,j} \log(\pi_{i,j}),$$

whose dual problem is

$$\min_{v \in \mathbb{R}^n} \left\{ G(v) = \frac{1}{n} \sum_{i=1}^{n} h_i(v) \right\}, \tag{EC.9}$$

where

$$h_i(v) = \eta \log \sum_j \exp\left(\frac{v_j - c_{i,j} - \eta}{\eta} \right) - \frac{1}{n} \sum_j v_j + \eta(1 + \log n).$$

Given the dual variable $v \in \mathbb{R}^n$, one can recover the primal variable π using

$$\pi(v) = \frac{\frac{1}{n} \exp\left(\frac{v_j - c_{i,j} - \eta}{\eta} \right)}{\sum_{j' \in [n]} \exp\left(\frac{v_{j'} - c_{i,j'} - \eta}{\eta} \right)}$$

Algorithm 2 essentailly optimizes the dual formulation (EC.9) with light computational speed.

THEOREM EC.3 **([75, Theorem 3])**. *Suppose we specify $T_{out} = \mathcal{O}(\frac{\|C\|_\infty \sqrt{\ln n}}{\epsilon}), T = n$, the number of total iterations (including outer and inner iterations) of Algorithm 2 is $\mathcal{O}(\frac{n\|C\|_\infty \sqrt{\ln n}}{\epsilon})$ with per-iteration cost $\mathcal{O}(n)$. Therefore, the number of arithmetic operations of Algorithm 2 for finding ϵ-optimal solution is $\mathcal{O}(\frac{n^2 \|C\|_\infty \sqrt{\ln n}}{\epsilon})$*

Algorithm 2 Stochastic Gradient-based Algorithm with Katyusha Momentum for solving OT [75]

Input: Accuracy $\epsilon > 0$, $\eta = \frac{\epsilon}{8 \log n}$, $\epsilon' = \frac{\epsilon}{6 \max_{i,j} c_{i,j}}$, maximum outer iteration T_{out}, and maximum inner iteration T.
Take $(y_0, z_0, \widetilde{\lambda}_0, \lambda_0, C_0, D_0) = (0, 0, 0, 0, 0, 0)$
for $t = 0, \ldots, T_{\text{out}} - 1$ **do**
  $\tau_{1,t} = \frac{2}{t+4}, \gamma_t = \frac{\eta}{9\tau_{1,t}}$
  $u_t = \nabla \phi(\widetilde{\lambda}_t)$
  for $j = 0, \ldots, T-1$ **do**
    $k = j + tT$
    $\lambda_{k+1} = \tau_{1,t} z_k + \frac{1}{2}\widetilde{\lambda}_t + (\frac{1}{2} - \tau_{1,t}) y_k$
    Sample i uniformly from $[n]$, and construct

$$H_{k+1} = u_t + \Big(\nabla h_i(\lambda_{k+1}) - \nabla h_i(\widetilde{\lambda}_t)\Big)$$

    Update $z_{k+1} = z_k - \gamma_t \cdot H_{k+1}/2$ and $y_{k+1} = \lambda_{k+1} - \eta H_{k+1}/9$
  end for
  Update $\widetilde{\lambda}_{t+1} = \frac{1}{T}\sum_{j=1}^{T} y_{tT+j}$
  Sample $\widehat{\lambda}_t$ uniformly from $\{\lambda_{tT+1}, \ldots, \lambda_{tT+T}\}$ and take $D_t = D_t + \texttt{vec}(\pi(\widehat{\lambda}_t))/\tau_{1,t}$
  $C_t = C_t + 1/\tau_{1,t}$
  $\pi_{t+1} = D_t / C_t$
end for
Query Algorithm 3 to Round $\widetilde{\pi} := \pi_{T_{\text{out}}}$ to $\widehat{\pi}$ such that $\widehat{\pi}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n$ and $\widehat{\pi}^{\mathrm{T}}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n$
Return $\widehat{\pi}$

Algorithm 3 Round to Γ_n ([1, Algorithm 2])

Input: $\pi \in \mathbb{R}_+^{n \times n}$
$X = \text{diag}(x_1, \ldots, x_n)$, with $x_i = \min\{1, \frac{1}{n r_i(\pi)}\}$. Here $r_i(\pi)$ denotes the i-th row sum of π.
$\pi' = X\pi$.
$Y = \text{diag}(y_1, \ldots, y_n)$, with $y_j = \min\{1, \frac{1}{n c_i(\pi')}\}$. Here $c_j(\pi')$ denotes the j-th column sum of π'.
$\pi'' = \pi' Y$.
$\mathbf{e}_r = \frac{1}{n}\mathbf{1}_n - r(\pi''), \mathbf{e}_c = \frac{1}{n}\mathbf{1}_n - c(\pi'')$, where

$$r(\pi'') = (r_i(\pi''))_{i \in [n]}, c(\pi'') = (c_j(\pi''))_{j \in [n]}.$$

Return $\pi'' + \mathbf{e}_r \mathbf{e}_c^{\mathrm{T}} / \|\mathbf{e}_r\|_1$.

EC.7. Proof of Lemmas 1, 2, and Theorem 4

Proof of Lemma 1. For the first part, it is noteworthy that $v(S)$ is associated with the objective

$$\widehat{F}(S)=\sum_{i,j}\widehat{\pi}_{i,j}\langle M_{i,j}^{\mathrm{T}}M_{i,j},S\rangle,$$

where $\widehat{\pi}_{i,j}$ is the ϵ-optimal solution to

$$F(S)=\min_{\pi\in\Gamma_n}\sum_{i,j}\pi_{i,j}\langle M_{i,j}^{\mathrm{T}}M_{i,j},S\rangle.$$

By definition, it holds that

$$0\le\widehat{F}(S)-F(S)\le\epsilon.$$

The second part follows from Theorem EC.3. □

The proof of Lemma 2 replies on the following technical result.

LEMMA EC.1 **([50])**. *Let $\{S_k\}_{k=1}^T$ be the updating trajectory of mirror ascent aiming to solve the maximization of $G(S)$ with $S\in\mathcal{S}_{2n}$, i.e.,*

$$S_{k+1}=\arg\max_{S\in\mathcal{S}_{2n}}\alpha\langle v(S_k),S\rangle+V(S,S_k),\quad k=1,\dots,T-1.$$

Here $v(S)$ is a supgradient of $G(S)$, and we assume there exists $M_*>0$ such that

$$\|v(S)\|_{Tr}:=\mathit{Trace}(v(S))\le M_*,\qquad\forall S\in\mathcal{S}_{2n}.$$

Let $\widehat{S}_{1:T}=\frac{1}{T}\sum_{k=1}^T S_k$, and S^* be a maximizer of $G(S)$. Define the diameter

$$D_{\mathcal{S}_{2n}}^2=\max_{S\in\mathcal{S}_{2n}}h(S)-\min_{S\in\mathcal{S}_{2n}}h(S)=\log(2n).$$

For constant step size

$$\alpha=\frac{D_{\mathcal{S}_{2n}}^2}{M_*\sqrt{T}}=\frac{\log(2n)}{M_*\sqrt{T}},$$

it holds that

$$0\le G(S^*)-G(\widehat{S}_{1:T})\le M_*\sqrt{\frac{4\log(2n)}{T}}.$$

Proof of Lemma 2. Let S^* and $\widehat{S}^*$ be maximizers of the objective $F(\cdot)$ and $\widehat{F}(\cdot)$, then we have the following error decomposition:

$$\begin{aligned}
&F(S^*)-F(\widehat{S}_{1:T})\\
=&\big[F(S^*)-\widehat{F}(S^*)\big]+\big[\widehat{F}(S^*)-\widehat{F}(\widehat{S}^*)\big]+\big[\widehat{F}(\widehat{S}^*)-\widehat{F}(\widehat{S}_{1:T})\big]+\big[\widehat{F}(\widehat{S}_{1:T})-F(\widehat{S}_{1:T})\big]\\
\le&2\epsilon+\big[\widehat{F}(S^*)-\widehat{F}(\widehat{S}^*)\big]+\big[\widehat{F}(\widehat{S}^*)-\widehat{F}(\widehat{S}_{1:T})\big]\\
\le&2\epsilon+\big[\widehat{F}(\widehat{S}^*)-\widehat{F}(\widehat{S}_{1:T})\big],
\end{aligned}$$

where the first inequality is because $\|F-\widehat{F}\|_\infty\le\epsilon$ and

$$|\big[F(S^*)-\widehat{F}(S^*)\big]|\le\epsilon,|\big[\widehat{F}(\widehat{S}_{1:T})-F(\widehat{S}_{1:T})\big]|\le\epsilon;$$

and the second inequality is because $\widehat{F}(\widehat{S}^*)-\widehat{F}(\widehat{S}_{1:T})\le 0$. It remains to bound $\big[\widehat{F}(\widehat{S}^*)-\widehat{F}(\widehat{S}_{1:T})\big]$. It is worth noting that

$$\|v(S)\|_{\mathrm{Tr}}=\sum_{i,j}\pi_{i,j}\|M_{i,j}M_{i,j}^{\mathrm{T}}\|_{\mathrm{Tr}}\le\sum_{i,j}\pi_{i,j}\cdot C=C.$$

Therefore, the proof can be finished by querying Lemma EC.1 with $M_*=C$ and stepsize $\alpha=\frac{\log(2n)}{C\sqrt{T}}$. □

Proof of Theorem 4. The proof can be finished by taking hyper-parameters such that

$$2\epsilon \le \frac{\delta}{2}, \qquad 2C\sqrt{\frac{\log(2n)}{T}} \le \frac{\delta}{2}.$$

In other words, we take $\epsilon = \frac{\delta}{4}$ and $T = \lceil \frac{16C^2 \log(2n)}{\delta^2} \rceil$. We follow the proof of Lemma 2 to take stepsize $\alpha = \frac{\log(2n)}{C\sqrt{T}}$. □

EC.8. Proof of Theorem 5

We reply on the following two technical results when proving Theorem 5.

THEOREM EC.4 **(Birkhoff-von Neumann Theorem [8])**. *Consider the discrete OT problem*

$$\min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} c_{i,j}, . \tag{EC.10}$$

There exists an optimal solution π that has exactly one entry of $1/n$ in each row and each column with all other entries 0.

THEOREM EC.5 **(Rank Bound, Adopted from [40, Theorem 2] and [39, Lemma 1])**. *Consider the domain set*

$$\mathcal{D} = \left\{ S \in \mathbb{S}_m^+ : \ \mathit{Trace}(S) = 1 \right\}$$

and the intersection of N linear inequalities:

$$\mathcal{E} = \left\{ S \in \mathbb{R}^{m \times m} : \ \langle S, A_i \rangle \ge b_i, i \in [N] \right\}.$$

Then, any feasible extreme point in $\mathcal{D} \cap \mathcal{E}$ has a rank at most $1 + \lceil \sqrt{2N + 9/4} - 3/2 \rceil$. Such a rank bound can be strengthened by replacing N by the number of binding constraints in $\mathcal{E}$.

Proof of Theorem 5. By taking the dual of inner OT problem, we find (SDR) can be reformulated as

$$\max_{\substack{S \in \mathcal{S}_{2n} \\ f,g \in \mathbb{R}^n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f_i + g_i) : \quad f_i + g_j \le \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \quad \forall i, j \in [n] \right\}. \tag{EC.11}$$

Let S^* be the optimal solution of variable S to the optimization problem above. Then for fixed S^*, according to Theorem EC.4 and complementary slackness of OT, there exists optimal solutions (f^*, g^*) such that only n constraints out of n^2 constraints in (EC.11) are binding. Moreover, an optimal solution to (SDR) can be obtained by finding a feasible solution to the following intersection of constraints:

$$\text{Find } S \in \mathcal{S}_{2n} \bigcap \mathcal{E} \triangleq \left\{ S : \ f_i^* + g_j^* \le \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \ \ i, j \in [n] \right\}.$$

By Theorem EC.5, any feasible extreme point from $\mathcal{S}_{2n} \cap \mathcal{E}$ has rank at most $1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$. Thus, we pick such a feasible extreme point to satisfy the requirement of Theorem 5.

Algorithm 4 Rank reduction algorithm for (SDR)

Run Algorithm 1 to obtain δ-optimal solution to (SDR), denoted as $\widehat{S}$.

`// Step 2: Find` n `binding constraints`

Run Hungarian algorithm [38] to solve OT (11) with $S \equiv \widehat{S}$, and obtain an optimal assignment $\sigma : [n] \to [n]$ together with dual optimal solutoin (f^*, g^*).

`// Step 3-9: Calibrate low-rank solution using a greedy algorithm`

Initialize $\delta^* = 1$

while $\delta^* > 0$ **do**

  Perform eigendecomposition $\widehat{S} = Q\Lambda Q^{\mathrm{T}}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ with $\mathrm{rank}(\widehat{S}) = r$

  Find a direction $Y = Q\Delta Q^{\mathrm{T}}$, where $\Delta \in \mathcal{S}^r$ is some nonzero matrix satisfying

$$\mathrm{Trace}(\Delta) = 0, \langle Q^{\mathrm{T}} M_{i,\sigma(i)} M_{i,j}^{\mathrm{T}} Q, \Delta \rangle = 0, \quad \forall i \in [n].$$

  If such Y does not exist, **then** break the loop.

  Take new solution $\widehat{S}(\delta^*) := \widehat{S} + \delta^* Y$, where

$$\delta^* = \arg\max_{\delta \geq 0} \left\{ \delta : \ \lambda_{\min}(\Lambda + \delta\Delta) \geq 0 \right\}.$$

  Update $\widehat{S} = \widehat{S}(\delta^*)$

end while

Return $\widehat{S}$

EC.9. Rank Reduction Algorithm

In this section, we develop a rank reduction algorithm that, based on the near-optimal solution (denoted as $\widehat{S}$) returned from Algorithm 1, finds an alternative solution of the same (or smaller) optimality gap while satisfying the desired rank bound in Theorem 5.

Step (i): Find n binding constraints. First, we fix $S \equiv \widehat{S}$ in (14) and finds the optimal solution (f^*, g^*) such that only n constraints out of n^2 constraints are binding. It suffices to apply the Huangarian algorithm [38] to solve the following balanced discrete OT problem

$$\max_{f,g \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n (f_i + g_i) : \ \ f_i + g_j \leq c_{i,j} \right\} = \min_{\pi \in \Gamma_n} \ \left\{ \sum_{i,j=1}^n \pi_{i,j} c_{i,j} \right\}$$

where $c_{i,j} = \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \widehat{S} \rangle$. The output of the Huangarian algorithm is a *deterministic* optimal transport that moves n probability mass points from the left marginal distribution of π to the right, which is denoted as a bijection σ that permutes $[n]$ to $[n]$. Thus, these n binding constraints are denoted as

$$f_i^* + g_{\sigma(i)}^* \leq \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \quad i \in [n].$$

We denote by the intersection of these n constraints as $\mathcal{E}_n$.

Step (ii): Calibrate low-rank solution using a greedy algorithm. Second, let us assume $\widehat{S}$ is not an extreme point of $\mathcal{S}_{2n} \cap \mathcal{E}_n$, since otherwise one can terminate the algorithm to output $\widehat{S}$ following the proof of Theorem 5. We run the following greedy rank reduction procedure:

(I) We find a direction $Y \neq 0$, along which $\widehat{S}$ remains to be feasible, and the null space of $\widehat{S}$ is non-decreasing.

(II) Then, we move $\widehat{S}$ along the direction Y until its smallest non-zero eigenvalue vanishes. We update $\widehat{S}$ to be such a new boundary point.

(III) We terminate the iteration until no movement is available.

To achieve (I), denote the eigendecomposition of $\widehat{S}$ with $\text{rank}(\widehat{S}) = r$ as

$$\widehat{S} = \begin{pmatrix} Q & 0 \end{pmatrix} \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q^{\mathrm{T}} & 0 \end{pmatrix} = Q \Lambda Q^{\mathrm{T}}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and $Q \in \mathbb{R}^{2n \times r}$. To ensure $\widehat{S} + \delta Y \in \mathcal{S}_{2n} \cap \mathcal{E}_n$ while $\text{Null}(\widehat{S} + \delta Y) \supseteq \text{Null}(\widehat{S})$, for some stepsize $\delta > 0$, it suffices to take

$$Y = \begin{pmatrix} Q & 0 \end{pmatrix} \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q^{\mathrm{T}} & 0 \end{pmatrix} = Q \Delta Q^{\mathrm{T}},$$

where $\Delta \in \mathcal{S}^r \setminus \{0\}$ is a symmetric matrix satisfying

$$\text{Trace}(\Delta) = 0, \quad \langle M_{i,j} M_{i,j}^{\mathrm{T}}, Q \Delta Q^{\mathrm{T}} \rangle = 0, \quad i \in [n].$$

To achieve (II), it suffices to solve the one-dimensional optimization

$$\delta^* = \arg\max_{\delta \geq 0} \Big\{ \delta : \ \lambda_{\min}(\Lambda + \delta \Delta) \geq 0 \Big\}, \tag{EC.12}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a given matrix. the optimization above admits closed-form solution update. Let eigenvalues of Δ be $\lambda'_1 \geq \cdots \geq \lambda'_r$. It suffices to solve

$$\delta^* = \arg\max_{\delta \geq 0} \Big\{ \delta : \ \min_{i \in [r]} \ (\lambda_i + \delta \lambda'_i) \geq 0 \Big\}.$$

As long as $\lambda'_r \geq 0$, we return $\delta^* = 0$. Otherwise, let i be an index such that $\lambda'_i \geq 0 > \lambda'_{i+1}$. We take $\delta^* = \max_{i < j \leq r} -\frac{\lambda_j}{\lambda'_j}$ as the desired optimal solution.

The overall algorithm is summarized in Algorithm 4. Its performance guarantee is summarized in Propositions EC.2, EC.3, and Theorem 6.

EC.10. Proof of Theorem 6

The proof of this theorem is separated into two parts.

PROPOSITION EC.2. *The rank of iteration points in Algorithm 4 strictly decreases. Thus, Algorithm 4 is guaranteed to terminate within $2n$ iterations.*

Proof. Assume on the contrary that $\text{rank}(\widehat{S}(\delta^*)) = \text{rank}(\widehat{S}) = r$. Since $\widehat{S}(\delta^*) = Q(\Lambda + \delta^*\Delta)Q^{\mathrm{T}}$, the positive eigenvalues of $\widehat{S}(\delta^*)$ are those of the matrix $\Lambda + \delta^*\Delta$. According to the solution structure of (EC.12), this could happen only when $\Lambda + \delta^*\Delta \succ 0$, i.e., either $\delta^* = 0$ or $\Delta \succeq 0$. For the first case, this algorithm terminates. For the second case, since $\text{Trace}(\Delta) = 0, \Delta \in \mathcal{S}^r$, it implies that $\Delta = 0$, which is a contradiction.

Thus, the rank of the iteration point reduces by at least 1 in each iteration. □

PROPOSITION EC.3. *Let S^* be the output of Algorithm 4. Then, it holds that*

(I) *S^* is a δ-optimal solution to (SDR).*

(II) *The rank of S^* satisfies*

$$\textit{rank}(S^*) \le 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor.$$

Proof. Recall the solution $\widehat{S}$ obtained from Step 1 of Algorithm 4 satisfies

$$\begin{aligned} F(\widehat{S}) &= \min_{\pi \in \Gamma_n} \sum_{i,j} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \widehat{S} \rangle \\ &= \max_{f,g \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n (f_i + g_i) : \ f_i + g_j \le \langle M_{i,j} M_{i,j}^{\mathrm{T}}, \widehat{S} \rangle \right\} \ge \texttt{objval}(SDR) - \delta. \end{aligned}$$

Since Step 2 of Algorithm 4 solves the OT problem exactly, we obtain

$$\frac{1}{n} \sum_{i=1}^n (f_i^* + g_i^*) = F(\widehat{S}) \ge \texttt{objval}(SDR) - \delta$$

Since Step 3-7 always finds feasible solutions to the n binding constraints

$$f_i^* + g_{\sigma(i)}^* \le \langle M_{i,j} M_{i,j}^{\mathrm{T}}, S \rangle, \quad i \in [n],$$

for any iteration points from Step 3-7, denoted as $\widetilde{S}$, the pair $(\widetilde{S}, f^*, g^*)$ is guaranteed to be the δ-optimal solution to (14), a reformulation of (SDR). Hence we finish the proof of Part (I).

For the second part, assume on the contrary that $r = \text{rank}(S^*) \ge 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$. It implies $n + 1 < r(r+1)/2$. Recall that Step 6 of Algorithm 4 essentially solves a linear system with $n+1$ constraints and $r(r+1)/2$ variables, so a nonzero matrix Δ is guaranteed to exist. Thus, one can pick a sufficiently small $\delta > 0$ such that $\lambda_{\min}(\Lambda + \delta\Delta) \ge 0$, which contradicts to the termination condition $\delta^* = 0$. Thus, we finish the proof of Part (II).

Combining both parts, we start to prove Theorem 6.

Proof. Algorithm 4 satisfies the requirement of Theorem 6. For computational complexity, the computational cost of Step 2 of Algorithm 4 is $\mathcal{O}(n^3)$. In each iteration from Step 3-7, the most computationally expansive part is to solve Step 6, which essentially solves a linear system with $n+1$ constraints and $r(r+1)/2$ variables. The conservative bound $r \le 2n$. Hence, the worst-case computational cost of Step 6 (which can be achieved using Gaussian elimination) is

$$\mathcal{O}((n + 1 + r(r+1)/2) \cdot (n+1)^2) = \mathcal{O}(n^4).$$

Since Algorithm 4 terminates within at most $2n$ iterations, the overall complexity of it is $\mathcal{O}(n^5)$.

EC.11. Numerical Implementation Details

When implementing our mirror ascent algorithm, for small sample size ($n \leq 50$), we use the exact algorithm adopted from https://pythonot.github.io/ to solve the inner OT; whears for large sample size we use the approximation algorithm adopt from https://github.com/YilingXie27/PDASGD to solve this subproblem.

For baseline approaches,

(I) the BCD method is implemented using the code from github.com/WalterBabyRudin/KPW_Test/tree/main;

(II) the max-sliced (MS) test is implemented using the same link;

(III) the mean embedding (ME) test is implemented using the code from https://github.com/wittawatj/interpretable-test;

(IV) the optimized MMD (MMD-O) test is implemented using the code from https://github.com/fengliu90/DK-for-TST.

We adopt from [70] to design KMS Wasserstein-based two-sample testing. For the synthetic dataset used in Fig. 2 and 3, we specify

$$\begin{aligned}
\mu &= \frac{1}{2}\mathcal{N}(0, 0.03 \cdot \mathbf{I}_D) + \frac{1}{2}\mathcal{N}(\mathbf{1}_D, 0.03 \cdot \mathbf{I}_D),\\
\nu &= \frac{1}{2}\mathcal{N}(0, \Sigma_1) + \frac{1}{2}\mathcal{N}(\mathbf{1}_D, \Sigma_2),\\
\text{where } \Sigma_1 &= \begin{pmatrix} 0.03 & -0.02 \\ -0.02 & 0.03 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.03 & 0.028 \\ 0.028 & 0.03 \end{pmatrix}.
\end{aligned}$$

For the synthetic dataset in Fig. 4, we specify

$$\begin{aligned}
\mu &= \frac{1}{2}\mathcal{N}(0, 0.03 \cdot \mathbf{I}_D) + \frac{1}{2}\mathcal{N}(\mathbf{1}_D, 0.03 \cdot \mathbf{I}_D),\\
\nu &= \frac{1}{2}\mathcal{N}(0, \Sigma_1) + \frac{1}{2}\mathcal{N}(\mathbf{1}_D, \Sigma_2),\\
\Sigma_1 &= 0.03 \cdot \mathbf{I}_D, \quad \Sigma_1[1,2] = \Sigma_1[2,1] \leftarrow -0.02\\
\Sigma_2 &= 0.03 \cdot \mathbf{I}_D, \quad \Sigma_2[1,2] = \Sigma_2[2,1] \leftarrow 0.028, \Sigma_2[1,1] = \Sigma_2[2,2] = 3.
\end{aligned} \tag{EC.13}$$

For the real dataset used in Fig. 2,

(I) When using Iris or credit dataset, we split its 70% as training set and 30% as testing set. We pre-process the data by doing the normalization (in 2-norm) for each feature vector. We compare the samples corresponding to labels 1 and 2 only;

(II) When using mnist dataset, we compare the samples corresponding to labels 6 and 8 only. We pre-process the data by passing through it into the sigmoid function such that all entries are bounded by $[0, 1]$.

EC.12. Additional Numerical Study

Statistical Convergence Rate. In this part, we validate the statistical rate of empirical KMS 2-Wasserstein distance (see the theoretical rate in Theorem 1) using the following configuration of blob dataset:

$$\begin{aligned}
\mu &= \frac{1}{2}\mathcal{N}(0, 0.03 \cdot \mathbf{I}_D) + \frac{1}{2}\mathcal{N}(\mathbf{1}_D, 0.03 \cdot \mathbf{I}_D), \\
\nu &= \frac{1}{2}\mathcal{N}(0, \Sigma_1) + \frac{1}{2}\mathcal{N}(\mathbf{1}_D, \Sigma_2), \\
\Sigma_1 &= 0.03 \cdot \mathbf{I}_D, \quad \Sigma_1[1,2] = \Sigma_1[2,1] \leftarrow -0.02 \\
\Sigma_2 &= 0.03 \cdot \mathbf{I}_D, \quad \Sigma_2[1,2] = \Sigma_2[2,1] \leftarrow 0.028, \Sigma_2[1,1] = \Sigma_2[2,2] = 4.
\end{aligned} \tag{EC.14}$$

In the left of Fig. EC.2, we plot the empirical KMS Wasserstein distance under H_0 for different sample size $n \in \{5, 10, 50, 100, 200, 500\}$ and dimension $D \in \{2, 5, 10, 20\}$, from which we see the statistic decays quickly to zero. In the right of Fig. EC.2, we plot the empirical KMS Wasserstein distance under H_1, from which we can see the statistic converges to a certain positive threshold. From both plots, we realize the convergence rate seems to be dimension-free, which makes the KMS distance a suitable choice for two-sample testing.

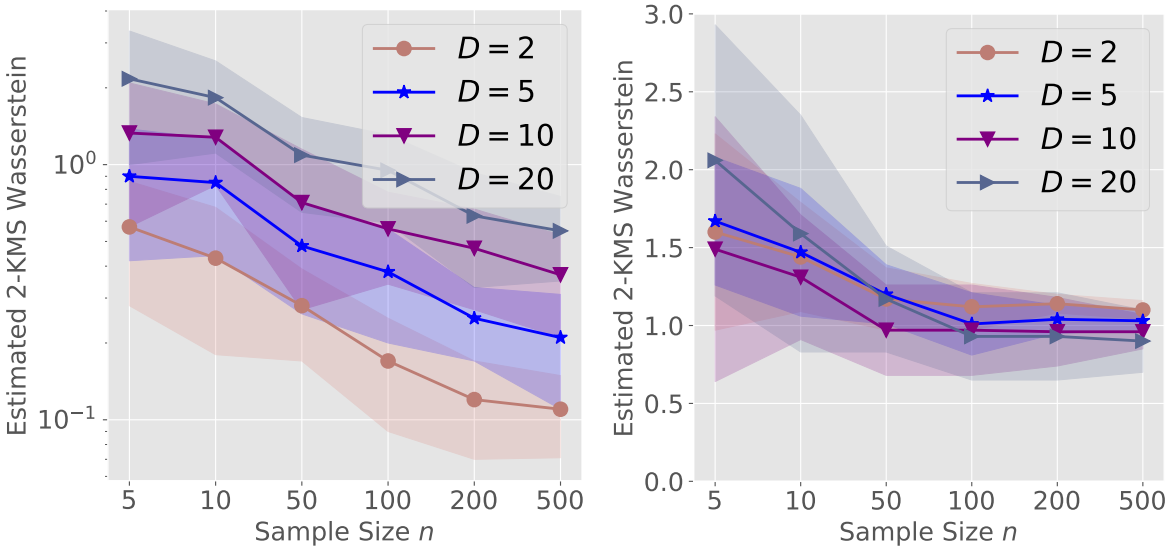

Figure EC.2 Value of empirical KMS Wasserstein distance for different choices of sample size and dimension. The error bar is generated using 20 independent trials.

Ablation Study on Kernel Choice and Bandwidth. Recall that we specified the kernel to be Gaussian in previous experiments, i.e., $K(x,y) = \exp(-\frac{\|x-y\|_2^2}{\sigma^2})$, where $\sigma^2 \equiv \widehat{\sigma}^2$ is picked using the median heuristic. In this part, we perform ablation study on two different choices of kernels: the Gaussian kernel and the exponential kernel $K(x,y) = \exp(-\frac{\|x-y\|_2}{\sigma})$. For Gaussian kernel, we validate the bandwidth choice

$$\sigma^2 \in \left\{ c \cdot \widehat{\sigma}^2 : \quad c \in \{0.1, 0.5, 1, 5, 10\} \right\},$$

and for exponential kernel, ew validate the bandwidth choice

$$\sigma \in \left\{ c \cdot \widehat{\sigma} : \quad c \in \{0.1, 0.5, 1, 5, 10\} \right\}.$$

The performance is examined based on the value of empirical KMS Wasserstein distance for data distributions defined in (EC.14) with $n = 500, d = 20$. Experiment results are reported in Fig. EC.3. From

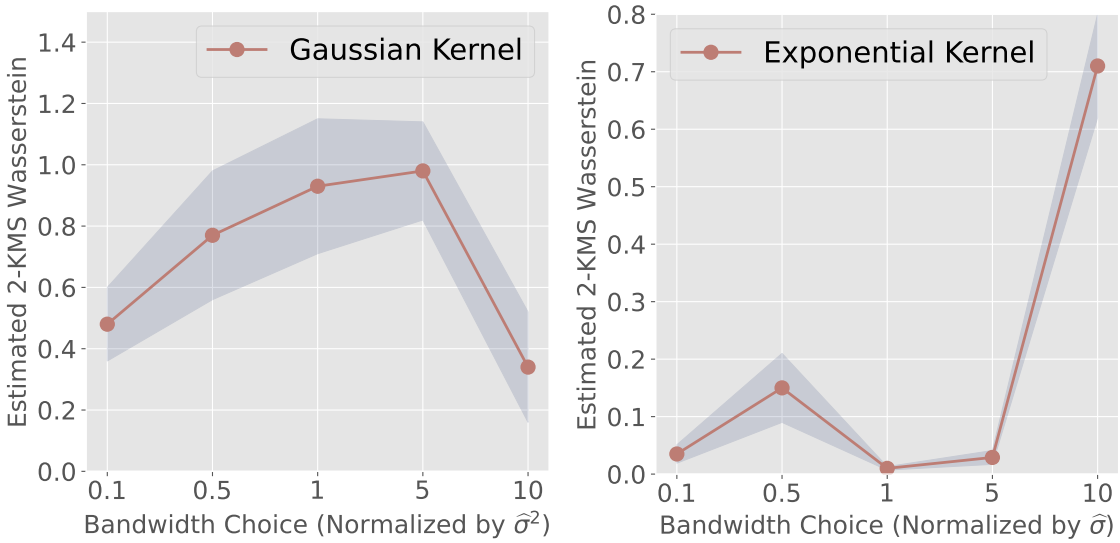

Figure EC.3 Ablation study for Gaussian and Exponential kernel choice.

the left plot, we realize that the bandwidth choice $1 \cdot \widehat{\sigma}^2$ achieves the near-optimal value for estimated KMS 2-Wasserstein distance, which justifies the median choice for Gaussian kernel is reasonable. On the other hand, if taking the exponential kernel, the bandwidth choice $10 \cdot \widehat{\sigma}$ achieves the largest value of estimated KMS 2-Wasserstein distance. Therefore, the median heuristic for the exponential kernel may not be optimal.

Details about Solution Rank. Recall that Theorem 5 provides the rank bound regarding some optimal solution from SDR. In this part, we compare the rank of the matrix estimated from Algorithm 1 with our theoretical rank bound based on the experiments from Fig. 2, which consists of four types of datasets (`blob`, `Iris`, `mnist`, and `credit`). For a given positive semidefinite matrix, we calculate the rank as the number of eigenvalues greater than the tolerance `1e-20`. We report the numerical performance on rank for these four datasets in Table EC.1 to EC.4, respectively.

It is noteworthy that for many instances, the optimal solution obtained from SDR has rank-1, which means in those cases, our SDR does not incur any optimality gap and solves Problem (9) exactly. Only for two cases in Table EC.4 are the rank from SDR higher than the rank bound. In these cases, one can run our rank reduction algorithm to obtain a low-rank solution.

Sample Size n	Rank Obtained from Algorithm 1	Rank Bound from Theorem 5
10	2	4
12	2	4
16	1	5
20	3	6
26	5	6
33	3	7
42	6	8
54	2	10

Table EC.1 Numerical performance on rank for `blob` dataset

Sample Size n	Rank Obtained from Algorithm 1	Rank Bound from Theorem 5
4	2	2
8	1	3
12	1	4
16	2	5
20	1	6
24	1	6
28	2	7
32	3	7

Table EC.2 Numerical performance on rank for `Iris` dataset

Sample Size n	Rank Obtained after Algorithm 1	Rank Bound from Theorem 5
50	2	9
100	3	9
150	5	16
200	3	19
250	2	21
300	3	24
350	1	26
400	2	27

Table EC.3 Numerical performance on rank for `credit` dataset

Sample Size n	**Rank Obtained from Algorithm 1**	**Rank Bound from Theorem 5**
50	1	9
100	2	13
150	4	16
200	3	19
250	7	21
300	**148**	24
350	3	26
400	**30**	27

Table EC.4 Numerical performance on rank for `mnist` dataset