# Minimax-Optimal Two-Sample Test with Sliced Wasserstein

**Binh Thuan Tran**
LAMA, LIGM, Université Gustave Eiffel
binh-thuan.tran3@univ-eiffel.fr

**Nicolas Schreuder**
CNRS, LIGM
nicolas.schreuder@cnrs.fr

## Abstract

We study the problem of nonparametric two-sample testing using the sliced Wasserstein (SW) distance. While prior theoretical and empirical work indicates that the SW distance offers a promising balance between strong statistical guarantees and computational efficiency, its theoretical foundations for hypothesis testing remain limited. We address this gap by proposing a permutation-based SW test and analyzing its performance. The test inherits finite-sample Type I error control from the permutation principle. Moreover, we establish non-asymptotic power bounds and show that the procedure achieves the minimax separation rate $n^{-1/2}$ with respect to the sliced Wasserstein distance over multinomial and bounded-support alternatives. This matches the optimal minimax rate $n^{-1/2}$ achieved by kernel-based tests with respect to the MMD, while leveraging the geometric structure of Wasserstein distances. Our analysis further quantifies the trade-off between the number of projections and statistical power. Finally, numerical experiments demonstrate that the test combines finite-sample validity with competitive power and scalability, and—unlike kernel-based tests, which require careful kernel tuning—it performs consistently well across all scenarios we consider.

## 1 INTRODUCTION

As machine-generated data becomes ubiquitous—from synthetic images to large language model outputs and scientific simulations—it is necessary to assess whether generated samples are statistically indistinguishable from real data samples. This question fits naturally into the framework of two-sample testing: one sample is drawn from a data distribution $\mu$ and the other from a generative model distribution $\nu$. Formally, the goal is to test

$$\mathcal{H}_0 : \mu = \nu \quad \text{against} \quad \mathcal{H}_1 : \mu \neq \nu.$$

Two-sample testing is a fundamental problem in statistics, with a long history and a wide range of practical applications (Lehmann and Romano, 2005). It has played a central role in areas such as clinical laboratory science (Miles et al., 2004; Zhang et al., 2021), finance (Horváth et al., 2014), bioinformatics (Borgwardt et al., 2006), neuroscience (Stelzer et al., 2013), and video content analysis (Liu et al., 2018). Recently, it has attracted growing interest in emerging machine learning and applied domains, including generative modeling (Li et al., 2017), model equality testing for large language models (Gao et al., 2024), and physics (Chakravarti et al., 2023).

Two primary classes of two-sample tests have been developed: *parametric* and *nonparametric*. Parametric tests, such as the classical $t$-test (Student, 1908) and Hotelling's two-sample $T^2$ test (Hotelling, 1931), rely on strong distributional assumptions (e.g., Gaussianity) and may lose power when these assumptions fail. Nonparametric tests avoid such assumptions and are thus more broadly applicable.

Among nonparametric methods, kernel-based tests have become especially prominent. In particular, the kernel two-sample test of Gretton et al. (2012), based on the Maximum Mean Discrepancy (MMD)–a special case of Integral Probability Metrics (Müller, 1997)–is widely used for its flexibility and computational tractability. However, kernel-based tests depend critically on kernel choice and hyperparameter tuning, which ideally should reflect the geometry of the underlying data (Schrab et al., 2023). This limitation motivates exploring alternatives such as optimal transport (OT), which incorporates geometric structure through

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

the choice of ground cost. The Wasserstein distance, in particular, is a natural candidate for two-sample testing, as it captures geometric discrepancies between distributions (Villani, 2008). However, the standard plug-in estimator suffers from the curse of dimensionality, with sharp error rates of order $n^{-1/d}$ (Fournier and Guillin, 2015; Weed and Bach, 2019).

To mitigate this curse of dimensionality, we focus on the sliced Wasserstein distance (Rabin et al., 2011), which projects measures onto one-dimensional subspaces and attains a parametric $n^{-1/2}$ convergence rate while retaining geometric interpretability (Nadjahi, 2021; Nietert et al., 2022). Despite its widespread use in generative modeling and learning (see Nguyen (2025, Chapter 6) for a summary), as well as promising empirical performance for testing (Grossi et al., 2025), theoretical guarantees for its application in two-sample testing remain largely unexplored. In particular, finite-sample validity, non-asymptotic power guarantees, and minimax optimality results for tests based on the sliced Wasserstein distances are currently lacking in the literature. This gap limits the theoretical understanding and rigorous adoption of sliced Wasserstein-based two-sample tests in practice.

In this paper, we address these gaps by proposing a permutation-based two-sample test using the sliced Wasserstein distance. We prove that this test enjoys finite-sample Type I error control. Furthermore, we establish non-asymptotic lower bounds on the test's power and demonstrate its minimax optimality against multinomial and bounded-support alternatives. We also analyze the computational–statistical trade-off inherent in the test, showing how the number of projections affects both runtime and statistical performance. A key technical contribution is our analysis of the SW distance between permuted distributions: to the best of our knowledge, this is the first work to control permutation quantiles for Wasserstein-based statistics. This analysis underpins the finite-sample guarantees for the SW test and extends naturally to the general Wasserstein distance. Finally, we complement our theoretical analysis with experiments, showing that our test achieves strong empirical performance and, unlike kernel-based methods whose power depends heavily on kernel choice, performs consistently well across diverse benchmarks without the need for parameter tuning or kernel aggregation.

The remainder of the paper is organized as follows. Section 2 reviews the necessary background on two-sample testing and the sliced Wasserstein distance. Section 3 discusses related work. Section 4 introduces the permutation-based sliced Wasserstein two-sample test. Section 5 presents our theoretical results on finite-sample validity, power analysis, and minimax optimality. Section 6 reports experimental evaluations on synthetic and real datasets. We conclude in Section 7 with a discussion of future research directions. All proofs are deferred to the Appendix.

## 2 BACKGROUND

In this section, we present the two-sample testing problem setup, define the sliced Wasserstein distance, and introduce minimax optimality for testing.

### 2.1 Problem setup

We consider two independent samples drawn from distributions $\mu$ and $\nu$ on $\mathbb{R}^d$:

$$\mathcal{Y}_n = (Y_1, \dots, Y_n) \overset{\text{i.i.d.}}{\sim} \mu, \quad \mathcal{Z}_m = (Z_1, \dots, Z_m) \overset{\text{i.i.d.}}{\sim} \nu,$$

both supported on a common ball $B_D = \{x \in \mathbb{R}^d : \|x\| \le D\}$ for some $D > 0$, and assume $n \le m$.

The goal is to test the hypotheses

$$H_0 : \mu = \nu \quad \text{versus} \quad H_1 : d(\mu, \nu) \ge \epsilon,$$

for some discrepancy metric $d$ and fixed $\epsilon > 0$. Many existing tests fit this framework by choosing $d$ as a specific probability discrepancy, such as the total variation distance or the MMD. Later, we will focus on the sliced Wasserstein distance.

A test is a measurable function

$$\phi_{n,m} : (\mathbb{R}^d)^n \times (\mathbb{R}^d)^m \to \{0, 1\},$$

where $\phi_{n,m} = 1$ indicates rejection of the hypothesis $H_0$. Typically, $\phi_{n,m}$ is based on a test statistic $\Psi$ measuring discrepancy between samples, rejecting $H_0$ if $\Psi$ exceeds a threshold.

A test's performance is characterized by its Type I (false positive) and Type II (false negative) errors, defined as follows. Let $\varepsilon > 0$ and define

$$\mathcal{P}_1(\varepsilon) \coloneqq \{(\mu, \nu) : d(\mu, \nu) \ge \epsilon\}.$$

A level-$\alpha$ test controls the Type I error uniformly by $\alpha \in (0, 1)$, and has power at least $1 - \beta$ if its Type II error is bounded by $\beta \in (0, 1)$ uniformly over $\mathcal{P}_1(\varepsilon)$:

$$\sup_{\mu=\nu} \mathbb{P}_{\mu\times\nu}(\phi_{n,m} = 1) \le \alpha,$$

$$\sup_{(\mu,\nu)\in\mathcal{P}_1(\varepsilon)} \mathbb{P}_{\mu\times\nu}(\phi_{n,m} = 0) \le \beta,$$

where probabilities are taken over the samples and any additional randomness (e.g., random projections).

In Section 5, we establish that our test is of level $\alpha$ and characterize the minimal separation between distributions detectable with power at least $1 - \beta$.

### 2.2 Sliced Wasserstein Distance

For $p \geq 1$, the $p$-Wasserstein distance (Villani, 2008, Section 6) between probability measures $\mu, \nu$ in the space of probability measures with finite $p$-th moments $\mathcal{P}_p(\mathbb{R}^d)$, is defined as

$$\mathrm{W}_p(\mu,\nu) := \left( \inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p \, d\pi(x,y) \right)^{1/p},$$

where $\Gamma(\mu,\nu)$ is the set of couplings with marginals $\mu$ and $\nu$. The $p$-Wasserstein distance $\mathrm{W}_p$ is a natural candidate for two-sample testing: it vanishes if and only if the distributions coincide and captures geometric discrepancies. However, its minimax separation rate scales as $n^{-c/d}$ for some $c > 1$, making it impractical in moderate to high dimensions (Ba et al., 2011; Chewi et al., 2025, Section 2.5).

To address this limitation, we use the sliced Wasserstein (SW) distance (Rabin et al., 2011), defined next.

**Definition 1 (SW distance)** *For $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the sliced Wasserstein distance is*

$$\mathrm{SW}_p(\mu,\nu) := \left( \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p \left( \Pi^\theta_\# \mu, \Pi^\theta_\# \nu \right) \sigma(d\theta) \right)^{1/p}, \tag{1}$$

*where $\Pi^\theta(x) = \langle \theta, x \rangle$ is the projection onto direction $\theta$ on the unit sphere $\mathbb{S}^{d-1}$, $\sigma$ is the uniform measure on the unit sphere, and we write $(T_\# \eta)$ for the push-forward measure associated to the probability measure $\eta \in \mathcal{P}_p(\mathbb{R}^d)$ and the measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$, i.e., $(T_\#\eta)(A) = \eta\big(T^{-1}(A)\big)$ for any measurable set $A$.*

By averaging one-dimensional Wasserstein distances over random projections, $\mathrm{SW}_p$ benefits from a dimension-free sample complexity rate while preserving geometric interpretability and has been successfully applied in large-scale statistical problems (see Sections 1 and 3). In order to estimate the SW distance, we replace the unknown measures $\mu$ and $\nu$ by their empirical counterpart

$$\widehat{\mu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{Y_i} \quad \text{and} \quad \widehat{\nu}_m := \frac{1}{m} \sum_{i=1}^m \delta_{Z_i}.$$

Moreover, we approximate the expectation with respect to the uniform distribution $\sigma$ on the sphere $\mathbb{S}^{d-1}$ by Monte Carlo sampling: we draw $L \geq 1$ independent directions $\Theta = (\theta_1, \ldots, \theta_L)$ from $\sigma$ and compute the tractable statistic

$$\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) := \frac{1}{L} \sum_{\ell=1}^L \mathrm{W}_p^p \left( \Pi^{\theta_\ell}_\# \widehat{\mu}_n, \Pi^{\theta_\ell}_\# \widehat{\nu}_m \right). \tag{2}$$

For brevity, we write $\widehat{\mathrm{SW}}_p^p$ for the estimator in (2).

### 2.3 Minimax optimality

Minimax theory provides a benchmark for testing by characterizing the smallest detectable difference for level-$\alpha$ tests (Ingster, 1993; Baraud, 2002). Define

$$\Phi_{n+m,\alpha} := \left\{ \phi_{n,m} : \sup_{\mu=\nu} \mathbb{P}_{\mu\times\nu}(\phi_{n,m} = 1) \leq \alpha \right\}$$

as the set of tests controlling Type I error at level $\alpha$ based on $n$ samples from $\mu$ and $m$ samples from $\nu$. Given $\epsilon > 0$, the minimax risk is then defined as the minimal worst-case Type II error over $\mathcal{P}_1(\epsilon)$ among all level-$\alpha$ tests,

$$R^\dagger_{n+m,\epsilon} := \inf_{\phi_{n,m} \in \Phi_{n+m,\alpha}} \sup_{(\mu,\nu)\in\mathcal{P}_1(\epsilon)} \mathbb{P}_{\mu\times\nu}(\phi_{n,m} = 0).$$

The minimax separation $\epsilon^\dagger_{n,m}$ is the smallest discrepancy $\epsilon > 0$ such that a level-$\alpha$ test exists with Type II error at most $\beta$:

$$\epsilon^\dagger_{n,m} := \inf \left\{ \epsilon > 0 : R^\dagger_{n+m,\epsilon} \leq \beta \right\}, \tag{3}$$

for fixed $\beta \in (0, 1-\alpha)$. Intuitively, $\epsilon^\dagger_{n,m}$ characterizes the detection boundary below which no $\alpha$-level test can reliably distinguish alternatives from the null.

For a given separation $\epsilon > 0$, the maximum Type II error of a test $\phi_{n,m} \in \Phi_{n+m,\alpha}$ over the class of alternative distributions $\mathcal{P}_1(\epsilon)$ is

$$R_{n+m,\epsilon}(\phi_{n,m}) := \sup_{(\mu,\nu)\in\mathcal{P}_1(\epsilon)} \mathbb{P}_{\mu\times\nu}(\phi_{n,m} = 0).$$

Then, its minimum separation is

$$\tilde{\epsilon}_{\phi,n,m} := \inf \left\{ \epsilon > 0 : R_{n+m,\epsilon}(\phi_{n,m}) \leq \beta \right\}.$$

A test $\phi_{n,m}$ is minimax rate-optimal if its minimum separation $\tilde{\epsilon}_{\phi,n,m}$ is equivalent to $\epsilon^\dagger_{n,m}$ up to constant factors. We show in Section 5 that the sliced Wasserstein test we consider is minimax rate-optimal.

## 3 RELATED WORK

The problem of two-sample testing has been extensively studied for over a century, with two main classes of approaches: *parametric* and *non-parametric*.

**Classical parametric tests.** Early approaches to the two-sample problem include the Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov, 1948), Student's $t$-test for mean comparison (Student, 1908), and Hotelling's $T^2$ test as a multivariate generalization of the $t$-test (Hotelling, 1931). These procedures are computationally simple and well

understood, but they are either restricted to the univariate setting (e.g., the KS test) or to testing specific moments such as the mean (e.g., $t$-test, Hotelling's $T^2$), which limits their applicability in high-dimensional or nonparametric problems.

**Kernel-based methods.** Kernel two-sample testing has emerged as a popular nonparametric alternative to classical parametric tests, designed to handle complex and high-dimensional data. The maximum mean discrepancy (MMD) test of Gretton et al. (2012) is widely used for its tractability and power, and it is closely related to energy distance (Székely et al., 2004) through the equivalence established by Sejdinovic et al. (2013). Because test performance is sensitive to kernel choice, later work has proposed kernel aggregation for adaptivity (Biggs et al., 2023; Schrab et al., 2023). An alternative strategy is data splitting (Sutherland et al., 2016), where a held-out subset is used to select the kernel and the remaining data are used to perform the test. As discussed in (Biggs et al., 2023, Section 1), this approach can reduce power because it decreases the effective sample size available for testing.

In parallel, scalable variants have been developed for large-scale problems. Examples include random Fourier feature approximations (Choi and Kim, 2024; Mukherjee and Sriperumbudur, 2025), Nyström subsampling (Chatalic et al., 2025) and coreset-based methods (Domingo-Enrich et al., 2023). While conceptually related to OT-based approaches—since both compare distributions via a discrepancy measure—kernel tests are grounded in reproducing kernel Hilbert space geometry, whereas OT-based methods exploit the geometry induced by transport costs.

**Optimal transport-based methods.** Parallel to kernel testing, there has been growing interest in tests based on optimal transport. Ramdas et al. (2017) survey connections between Wasserstein distance, MMD, and energy distance, framing entropic OT as interpolating between transport and kernel discrepancies. Building on this, several works analyze the statistical properties of Wasserstein-based tests (Imaizumi et al., 2022; González-Delgado et al., 2023) and propose projection-based approaches such as projected Wasserstein tests (Wang et al., 2021, 2022). Sliced variants, which exploit the tractability of one-dimensional OT, have been studied extensively: the sliced Wasserstein distance was introduced by Rabin et al. (2011), with theoretical properties developed by Bonnotte (2013). Generalizations include generalized and max-sliced Wasserstein distances (Kolouri et al., 2019; Deshpande et al., 2019), subspace-robust Wasserstein (Paty and Cuturi, 2019), and recent work on statistical guarantees for max-sliced variants (Boedihardjo, 2025; Wang et al., 2025). Closest to our work, Hu and Lin (2025) propose a max-sliced Wasserstein test with bootstrap calibration and asymptotic validity, while finite-sample guarantees remain an open question.

**Project–then–test methods.** A related line of work uses random projections to reduce dimensionality before applying classical test statistics. For example, Lopes et al. (2011) propose projecting high-dimensional data onto random directions and then applying Hotelling's $T^2$ test.

**Calibration and optimality.** Permutation testing plays a central role in our approach. Its finite-sample validity has been long established (Hoeffding, 1952; Hemerik and Goeman, 2018), and more recent results show that permutation tests can achieve the minimax optimal detection boundary (Kim et al., 2022).

In summary, classical tests are restricted in scope, kernel tests provide flexible and efficient nonparametric alternatives, and OT-based tests capture geometric aspects of the distributions but often lack finite-sample guarantees. We propose a permutation-based sliced Wasserstein test that combines size control with non-asymptotic power bounds, achieves the minimax rate $n^{-1/2}$, and quantifies the trade-off between computational efficiency and statistical power.

# 4 PERMUTATION SW TEST

Before presenting our method in Algorithm 1, we detail the permutation procedure and the obtained test.

## 4.1 Permutation procedure

Recall $\mathcal{Y}_n = (Y_1, \dots, Y_n)$ and $\mathcal{Z}_m = (Z_1, \dots, Z_m)$ denote the two samples. Let $\Theta = (\theta_1, \dots, \theta_L)$ be the set of projection directions (drawn i.i.d. uniformly from the sphere). We form the pooled dataset $\mathcal{X}_{n+m} = (X_1, \dots, X_{n+m})$ as

$$X_i \coloneqq Y_i \text{ for } 1 \le i \le n,\ X_{n+i} \coloneqq Z_i \quad \text{for } 1 \le i \le m.$$

For a permutation $\pi \in S_{n+m}$, the symmetric group on $\{1, \dots, n+m\}$, define the permuted dataset $\mathcal{X}^{\pi}_{n+m} = (X_{\pi(1)}, \dots, X_{\pi(n+m)})$. Construct empirical measures

$$\widehat{\mu}^{\pi}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_{\pi(i)}}, \quad \widehat{\nu}^{\pi}_m \coloneqq \frac{1}{m}\sum_{i=n+1}^{n+m} \delta_{X_{\pi(i)}},$$

and compute the empirical sliced Wasserstein distance

$$\widehat{\mathrm{SW}}^{p,\pi}_p \coloneqq \widehat{\mathrm{SW}}^{p}_p(\widehat{\mu}^{\pi}_n, \widehat{\nu}^{\pi}_m). \tag{4}$$

Let $N := n+m$, and denote by $F_{\widehat{\mathrm{SW}}_p^{p,\pi}}$ the permutation empirical cumulative distribution function:

$$F_{\widehat{\mathrm{SW}}_p^{p,\pi}}(t) := \frac{1}{|S_N|} \sum_{\pi \in S_N} \mathbf{1}\{\widehat{\mathrm{SW}}_p^{p,\pi} \leq t\},$$

where $S_N$ is the set of all permutations of $\{1, \ldots, N\}$. We write the $1-\alpha$ quantile of this distribution as

$$c_{1-\alpha,N} := \inf\{t : F_{\widehat{\mathrm{SW}}_p^{p,\pi}}(t) \geq 1-\alpha\}. \tag{5}$$

**Decision making.** Given the quantile $c_{1-\alpha,N}$, we reject the null hypothesis if $\widehat{\mathrm{SW}}_p^p > c_{1-\alpha,N}$. This critical value ensures finite-sample Type I error control under the permutation-invariance (exchangeability) assumption, which holds for both two-sample and independence testing problems (see Section 5.1).

**Remark 2** *Exact computation of the critical value* (5) *is generally infeasible for large samples, so it is commonly approximated via Monte Carlo simulations. The relationship between the exact critical value and its Monte Carlo approximation is discussed in Domingo-Enrich et al. (2023, Lemma 6).*

Let $r$ be the uniform distribution on $S_N$ and let $\mathbb{Z}_B := (\pi_b)_{1 \leq b \leq B}$ be a collection of i.i.d. samples from $r$. Define $\pi_{B+1} :=$ id as the identity permutation. Following the approach of Romano and Wolf (2005, Lemma 1), and in order to achieve the prescribed non-asymptotic test level, we then consider the statistics $\widehat{\mathrm{SW}}_p^{p,\pi_b}$ for $b = 1, \ldots, B+1$ and estimate the $(1-\alpha)$ quantile of the permutation distribution by

$$\widehat{c}^B_{1-\alpha,N} := \inf\left\{t : \frac{1}{B+1} \sum_{i=1}^{B+1} \mathbf{1}\{\widehat{\mathrm{SW}}_p^{p,\pi_i} \leq t\} \geq 1-\alpha\right\}.$$

Letting $\widehat{\mathrm{SW}}_p^{p,r_{\bullet 1}} \leq \widehat{\mathrm{SW}}_p^{p,r_{\bullet 2}} \leq \cdots \leq \widehat{\mathrm{SW}}_p^{p,r_{\bullet (B+1)}}$ denote the order statistics of $\{\widehat{\mathrm{SW}}_p^{p,\pi_b}\}_{b=1}^{B+1}$, the quantile estimator equals $\widehat{\mathrm{SW}}_p^{p,r_{\bullet \lceil (B+1)(1-\alpha) \rceil}}$. The Sliced Wasserstein Test is then defined as

$$\Delta(\mathcal{Y}_n, \mathcal{Z}_m, \mathbb{Z}_B, \Theta) := \mathbf{1}\big(\widehat{\mathrm{SW}}_p^p > \widehat{c}^B_{1-\alpha,N}\big).$$

Algorithm 1 summarizes the testing procedure.

**Remark 3** *While bootstrap or subsampling methods are often used to calibrate critical values (see, e.g., Hu and Lin (2025)), their asymptotic validity rely on well-behaved limiting distributions and do not guarantee uniform finite-sample size control, as required for the finite sample minimax framework that we consider.*

**Algorithm 1** SW-Permutation Test

**Require:** Datasets $\mathcal{Y}_n = (Y_1, \ldots, Y_n)$ and $\mathcal{Z}_m = (Z_1, \ldots, Z_m)$; significance level $\alpha \in (0,1)$; number of permutations $B$; number of projections $L$.

**Ensure:** Decision $\Delta \in \{0,1\}$

1: Form pooled dataset $\mathcal{X}_{n+m} = (X_1, \ldots, X_{n+m})$ as

$$X_i := \begin{cases} Y_i & 1 \leq i \leq n, \\ Z_{i-n} & n < i \leq n+m \end{cases}$$

2: Sample $\theta_1, \ldots, \theta_L \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$

3: Sample $\pi_1, \ldots, \pi_B \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(S_{n+m})$

4: Set $\pi_{B+1} :=$ id (identity permutation)

5: **for** $b = 1$ to $B+1$ **do**

6: Define empirical measures

$$\widehat{\mu}_n^{\pi_b} := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_{\pi_b(i)}}, \quad \widehat{\nu}_m^{\pi_b} := \frac{1}{m} \sum_{i=n+1}^{n+m} \delta_{X_{\pi_b(i)}}$$

7: Compute statistic $\widehat{\mathrm{SW}}_p^{p,\pi_b} := \widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n^{\pi_b}, \widehat{\nu}_m^{\pi_b})$

8: **end for**

9: Set the critical $\widehat{c}^B_{1-\alpha,n+m}$ as the $(1-\alpha)$ empirical quantile of $\{\widehat{\mathrm{SW}}_p^{p,\pi_b}\}_{b=1}^{B+1}$

10: **return** $\Delta := \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_{B+1}} > \widehat{c}^B_{1-\alpha,n+m}\right)$

### 4.2 Computational complexity

With a one-time pre-computation of projections $O(LdN)$ and per-projection sorting $O(LN\log(N))$, and the main loop $O(LBN)$, our method achieves an overall time complexity of $O\big(LN(d + \log N + B)\big)$. Moreover, storing the original and projected datasets, as well as the permuted statistics requires space complexity $O(Nd + NL + B)$.

Note that, choosing $L = N$ projections leads to quadratic time complexity (up to logarithmic factors) in the sample size, comparable to that of the MMD two-sample test. Reducing the number of projections can mitigate this cost, trading off between statistical power and computational efficiency. A characterization of this trade-off can be obtained from Theorem 5.

## 5 THEORETICAL ANALYSIS

In this section, we analyze the proposed test by studying its level and power. Without loss of generality, we assume $n \leq m$. We let $B \geq 1$ denote the number of permutations and $L \geq 1$ the number of projections. For space constraints and to improve readability, all proofs are deferred to the Appendix and all constants

are made explicit in the proof.

### 5.1 Level of the test

We begin by establishing that the proposed test controls the Type I error non-asymptotically.

It is now a well-known fact that permutation-based tests control the Type I error at finite samples for any test statistic when the data are exchangeable under the null hypothesis (Hoeffding, 1952; Lehmann and Romano, 2005). This property is crucial since it ensures validity without asymptotic approximations. In the two-sample setting, this holds under $H_0 : \mu = \nu$ since all observations are i.i.d. and the joint distribution is invariant under permutations of group labels. We formalize this below.

**Theorem 4 (Type I error Control)** *The test $\Delta$ defined in Algorithm 1 has non-asymptotic level $\alpha$ for any $\alpha \in (0, 1)$. That is*

$$\mathbb{P}_{\mu\times\mu\times r\times\sigma}\left(\Delta\left(\mathcal{Y}_n, \mathcal{Z}_m, \mathbb{Z}_B, \Theta\right) = 1\right) \leq \alpha,$$

*where the probability is taken over the samples, projection directions, and permutation randomness.*

### 5.2 Power of the test

While Type I error control follows from standard permutation arguments, analyzing the test's power is more challenging. The main difficulty is that the critical value is defined as a data-dependent permutation quantile of the sliced Wasserstein statistic. For MMD, existing analyses exploit its $U$- or $V$-statistic structure together with associated concentration inequalities (Schrab et al., 2023; Domingo-Enrich et al., 2023), but the sliced Wasserstein distance does not admit such a representation, preventing the direct use of these techniques. To address this issue, we leverage properties of the sliced Wasserstein distance under permutations, in particular, a permutation bounded-differences (McDiarmid-type) inequality and an optimal matching bound from Bobkov and Ledoux (2021, Corollary 5). These ingredients yield the non-asymptotic power bound stated in Theorem 5.

**Theorem 5 (Power Control)** *Let $\beta \in (0, 1)$ and $\frac{1}{B+1} \leq \alpha < 1$. The test described in Algorithm 1 has power at least $1 - \beta$ provided that*

$$\mathrm{SW}_p^p(\mu, \nu) \geq \frac{C(D, p, \alpha, \beta, B)}{\sqrt{L \wedge n}},$$

*where the constant $C(D, p, \alpha, \beta, B)$ is made explicit in the proof.*

This result shows that the SW test achieves the parametric separation rate $n^{-1/2}$ up to constants. To match the sample size, the number of projections $L$ should scale proportionally to $n$, ensuring that projection variance does not limit power.

### 5.3 Minimax optimality

To assess the statistical optimality of the proposed test, we now study its minimax separation rate. Establishing minimax lower bounds in nonparametric testing typically involves constructing least favorable alternatives and adapting classical arguments, which is the approach we take here. These are the first minimax lower bounds for two-sample testing with respect to the sliced Wasserstein distance. Consider the class

$$\mathcal{P}_{\mathbb{R}^d}(D) := \left\{\mu \in \mathcal{P}(\mathbb{R}^d) : \mathrm{diam}(\mathrm{supp}(\mu)) \leq D\right\},$$

comprising all probability measures on $\mathbb{R}^d$ with support diameter at most $D > 0$. Taking $L \geq cn$ for some absolute constant $c > 0$, Theorem 5 implies an upper bound on the minimax separation rate (see Eq. (3))

$$\epsilon^{\dagger}_{n,m} \leq Cn^{-1/2}.$$

We now establish lower bounds in two scenarios to assess the tightness of the upper bound. First, we provide a lower bound for the simpler case of multinomial distributions. We will then consider more generally distributions with bounded support. The following result is obtained via an adaptation of the classical lower bound technique from Ingster (1987).

**Proposition 6** *Let $\alpha \in \left(\frac{1}{B+1}, 1\right)$ and $\beta \in (0, 1 - \alpha)$. Given an integer $d \geq 1$, define $[d] := \{1, \ldots, d\}$, and let $\mathcal{P}^{(d)}_{\mathrm{Multi}}$ denote the class of multinomial distributions on $[d]$. Then the minimax separation rate of the two-sample testing problem over $\mathcal{P}^{(d)}_{\mathrm{Multi}} \times \mathcal{P}^{(d)}_{\mathrm{Multi}}$ satisfies*

$$\epsilon^{\dagger}_{n,m} \geq \frac{C(d, \alpha, \beta)}{\sqrt{n}},$$

*where the constant $C(d, \alpha, \beta)$ is made explicit in the proof.*

The above result establishes that no test can achieve faster than $n^{-1/2}$ separation in the simple multinomial setting.

Turning to the more general class of distributions with bounded support $\mathcal{P}_{\mathbb{R}^d}(D)$, we obtain the following lower bound using Le Cam's two-point method (Le Cam, 1973, 2012).

**Proposition 7** *Let $\alpha, \beta \in (0, 1)$ with $\alpha + \beta < 0.5$. For the two-sample testing problem over $\mathcal{P}_{\mathbb{R}^d}(D)$, the minimax separation rate satisfies*

$$\epsilon^{\dagger}_{n,m} \geq \frac{C(D, \alpha, \beta, p)}{\sqrt{n}}.$$

Thus, the lower bound extends beyond discrete settings to general bounded-support distributions.

Together with Theorem 5, these results imply that the permutation SW test achieves the minimax optimal separation rate $n^{-1/2}$ over both multinomial and bounded-support distributions, matching known optimal rates for MMD-based tests (Kim et al., 2022, Proposition 4.4; Kim and Schrab, 2023, Theorems 7,8).

# 6 NUMERICAL EXPERIMENTS

To complement our theoretical findings, we empirically evaluate the performance of the SW test described in Algorithm 1. We focus on the balanced case $n = m$ and assess performance across various numbers of projections $L$. We compare against several baselines: the Projected Wasserstein (PW) test (Wang et al., 2021) and the MMD test (Gretton et al., 2012). For MMD, we consider linear, Gaussian, and Laplace kernels, with bandwidths for the latter two selected via the median heuristic (Garreau et al., 2017). For the PW test, we use the authors' official implementation with default settings (projection dimension 3 and a 50/50 split for training and testing), as no practical guidance on these parameters was provided. All experiments were run on a machine with an Intel Core Ultra 9 185H CPU and 64 GB of RAM.

As discussed in Section 3, several scalable variants of MMD have been developed for large-scale problems. However, these approximations do not increase statistical power beyond standard MMD (see, e.g., Choi and Kim (2024, Section 5)), so we exclude them from our benchmarks and focus on contrasting how SW- and MMD-based tests distinguish distributions. Moreover, we also exclude MMD aggregation methods: while aggregation improves adaptivity and can outperform individual kernels, our goal is to compare discrepancies rather than kernel selection strategies, so we restrict attention to standard kernels.

In all experiments, we set the significance level to $\alpha = 0.05$ and use $B = 200$ permutations. Empirical power is estimated from 150 independent repetitions, parallelized with Joblib. Error bars represent 95% confidence intervals computed from the normal approximation of the empirical proportion (mean $\pm 1.96$ standard errors).

## 6.1 Synthetic datasets

Our first experiment evaluates power under a Gaussian covariance shift: samples are drawn from $\mu = \mathcal{N}(0, I_{60})$ and $\nu = \mathcal{N}(0, \Sigma)$, where $\Sigma = \mathrm{diag}(\delta^2, \delta^2, 1, \ldots, 1)$ with shift magnitude $\delta = 2.7$.

The second scenario examines geometric differences between distributions: the uniform distribution on the 5D unit sphere versus the uniform distribution on the 5D unit ball. This setting tests performance when one distribution lies on a lower-dimensional manifold embedded within the other's support.

As shown in Figure 1, the PW test achieves the highest power under covariance shifts, consistent with Wang et al. (2022), while the SW test performs competitively with Gaussian and Laplace MMD.

In the ball–vs.–sphere scenario, the MMD test with a Laplace kernel attains the best performance, but the SW test substantially outperforms both Gaussian MMD and PW, even with a limited number of projection directions. The weaker performance of Gaussian MMD likely stems from over-smoothing under the median heuristic, which blurs sharp radial differences, whereas the Laplace kernel's slower decay preserves more of this contrast. By contrast, MMD with a linear kernel performs worst across both settings, since with $k(x, y) = x^\top y$ it reduces to a mean-difference test, which vanishes in these cases. Overall, while the SW test is not always the most powerful, it delivers consistently strong performance across benchmarks without requiring parameter tuning, demonstrating robustness to diverse distributional differences.

## 6.2 Benchmark datasets

To complement the synthetic experiments, we further evaluate the proposed test on four benchmark datasets commonly used in the machine learning literature. These experiments serve two purposes. First, on the handwritten digits dataset, we investigate the behavior of the test across a range of sample sizes in order to assess how its empirical power evolves with $n$. Second, on three additional benchmark datasets, namely Moon, EMNIST, and SUSY, we report results for one representative sample size in order to emphasize the practical value of the proposed procedure on structured real or standard benchmark data.

**Digits dataset.** We evaluate the two-sample testing procedures on the handwritten digits dataset from `scikit-learn`. Let $\mu_6$ denote the uniform distribution over images of the digit 6 and $\mu_9$ the corresponding distribution for digit 9. We define the mixture $\nu = 0.85\mu_6 + 0.15\mu_9$. Following Wang et al. (2022), we preprocess the dataset by applying a sigmoid transformation to each image so that all pixel values lie within $[0, 1]$. The SW test effectively discriminates between $\mu_6$ and $\nu$ by capturing localized intensity differences in 1D projections. As illustrated in Figure 1, this experiment highlights the SW test's ability to detect subtle distributional differences in high-dimensional image

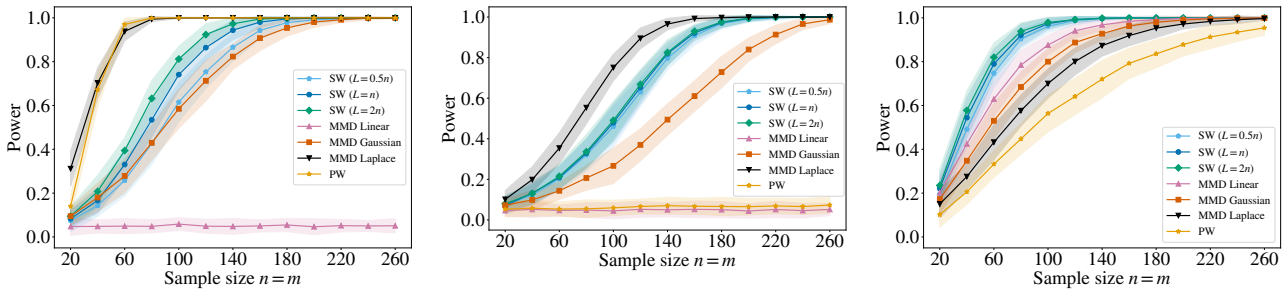


Figure 1: Power across three scenarios: Gaussian covariance shift, ball vs. sphere, and Digits mixture.

data without requiring kernel parameter tuning, confirming its practical applicability.

We next consider three additional datasets and report empirical power for one representative choice of sample size in each case, in order to further illustrate the practical effectiveness of the proposed test.

**Moon dataset.** We use the Moon dataset generated by the `make_moons` function in the `scikit-learn` library, which consists of two interleaving crescent-shaped clusters. In our two-sample testing setup, $\mu$ denotes the distribution supported on class 1 only. The alternative distribution $\nu$ is defined as a perturbed mixture: each observation is drawn from class 1 with probability 0.85 and from class 0 with probability 0.15.

**EMNIST dataset.** Inspired by Domingo-Enrich et al. (2023), we conduct experiments on the EMNIST dataset introduced by Cohen et al. (2017), which is an extension of MNIST by LeCun et al. (2012). In our two-sample testing setup, we use the training set of the *Letters* split and focus on the visually similar characters M (label 13) and N (label 14). We define $\mu$ as the distribution supported on class M only, and construct $\nu$ as a perturbed mixture: each observation is drawn from class M with probability 0.85 and from class N with probability 0.15.

**SUSY Dataset.** Following Chatalic et al. (2025), the last experiment is conducted on the Susy dataset, originally presented in Baldi et al. (2014), which consists of Monte Carlo simulations of high-energy physics collider events, containing two classes: background and signal. The dataset includes 18 features, where the last 10 are deterministic functions of the first 8; these are referred to as high-level and low-level features, respectively. Following the experimental setup of Chatalic et al. (2025), we restrict our analysis to the 8 low-level features. In our two-sample testing scenario, distribution $\mu$ is composed exclusively of background events, while $\nu$ is constructed as a mixture of background and signal events, with each sample drawn from the signal class with probability 15%, and from the background class with probability 85%.

For the last three datasets, the empirical power results are reported in Table 1, with standard deviations shown in parentheses. Overall, these results further support the practical relevance of the proposed test across datasets of different dimensions and structures.

### 6.3 Effect of the Number of Projections

While the SW test generally performs well on our three benchmark datasets, it underperforms relative to MMD tests with Gaussian, Laplace, and linear kernels in an additional Gaussian mean-shift experiment for a range of projection numbers $L \in \{0.5n, n, 2n\}$ (see Section 15). To better understand this behavior, we examine the effect of increasing the number of projections $L$ on both power and computation time. Fixing $n = m = 140$ in the mean-shift setting, Figure 2 shows that power improves steadily with larger $L$, though at the cost of increased computation time (Figure 4).

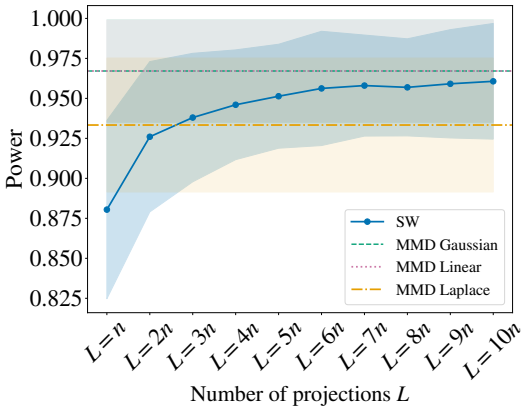


Figure 2: SW Test power vs. number of projections (fixed sample sizes $n = m = 140$)

Fortunately, the computational bottleneck of permutation-based SW testing is mitigated by its projection structure: since it computes many

Table 1: Power on the last three datasets in Section 6.2, with standard deviations reported in parentheses.

| **Dataset** | **MMD Gaussian** | **MMD Laplace** | **MMD Linear** | **SW** ($L = 0.5n$) | **SW** ($L = n$) |
|---|---|---|---|---|---|
| Moons ($d = 2$, $n = m = 300$) | 0.845 (0.026) | 0.89 (0.024) | 0.898 (0.022) | **0.999** (0.002) | **0.999** (0.002) |
| EMNIST ($d = 784$, $n = m = 400$) | 0.929 (0.022) | 0.913 (0.021) | 0.933 (0.019) | 0.955 (0.019) | **0.960** (0.017) |
| SUSY ($d = 8$, $n = m = 800$) | 0.432 (0.034) | 0.593 (0.042) | 0.563 (0.032) | **0.729** (0.028) | **0.729** (0.026) |

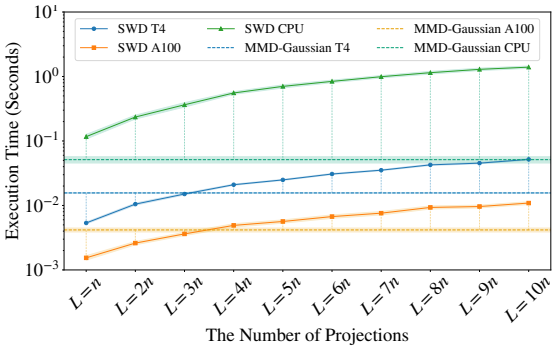


Figure 3: Computation time (log scale) of SW tests

independent 1D projections, the workload is naturally amenable to GPU parallelization. This allows scaling the number of projections with only moderate increases in wall-clock time, given sufficient GPU resources. In contrast, MMD computations hinge on the kernel matrix, whose size is fixed once the sample size is given, limiting further scaling through additional parallelism. In practice, we find that SW implementations can better exploit hardware parallelism and thus often scale better than naive MMD implementations under similar resource constraints.

To validate this, we compared computation times using two sets of 140 samples from a 60-dimensional mean-shifted Gaussian distribution. We measured the average execution time over 100 repetitions on a local CPU, Google Colab's T4, and A100 GPUs. Results in Figure 3 confirm substantial speedups for the SW test on GPUs, though larger-scale studies would be valuable to fully assess its scalability in practice. Finally, to illustrate Theorem 4, we assessed the empirical Type I error of the SW test. In each of 2000 repetitions, two independent samples of size 50 were drawn from $\mathcal{N}(0, I_{60})$. The results, reported in Table 2 for three choices of $L$, show that the permutation-based SW test effectively controls the Type I error rate even at small sample sizes.

[1]To leverage available CPU resources, we perform independent repetitions in parallel across all CPU cores and report the mean wall-clock time per repetition. Due to heterogeneous core performance, the average time can exceed that observed when using a single core. All methods are timed on the same inputs within each repetition, so their relative ranking is unaffected.

Table 2: Type I error for the SW test

| **Test Statistics** | **Type I Error** |
|---|---|
| SW ($L = 0.5n$) | $0.04982 \pm 0.00543$ |
| SW ($L = n$) | $0.04844 \pm 0.00501$ |
| SW ($L = 2n$) | $0.04961 \pm 0.00461$ |

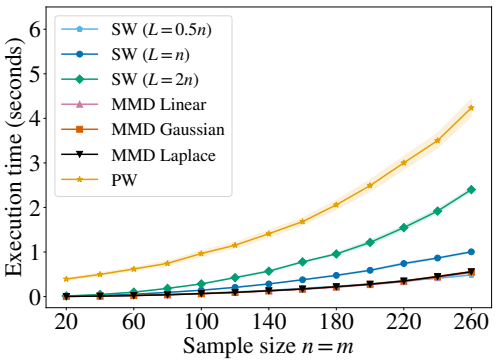


Figure 4: Computation time for ball vs. sphere [1]

In summary, our experiments demonstrate that the SW test offers a robust, parameter-free alternative to kernel-based methods, balancing statistical power and computational efficiency, especially when leveraging parallel hardware.

## 7 CONCLUSION

We proposed a permutation-based two-sample test using the sliced Wasserstein distance, establishing finite-sample validity, non-asymptotic power bounds, and minimax optimality over multinomial and bounded-support alternatives. To our knowledge, this is the first Wasserstein-based test with finite-sample guarantees. Our analysis quantified the trade-off between the number of projections and statistical power, and experiments showed that the test achieves consistently strong performance across benchmarks without parameter tuning, while remaining scalable on parallel hardware. Future directions include extending minimax optimality results to broader smoothness classes and exploring adaptive or non-uniform slicing strategies (e.g., generalized projections (Kolouri et al., 2019)).

**Acknowledgements**

The authors would like to thank Pierre Bizeul for pointing out an interesting reference and Mohamed Hebiri for helpful discussions.

## References

Albert, M. (2015). *Tests of independence by bootstrap and permutation: an asymptotic and non-asymptotic study. Application to neurosciences.* PhD thesis, Université Nice Sophia Antipolis.

Albert, M. (2019). Concentration inequalities for randomly permuted sums. In *High Dimensional Probability VIII: The Oaxaca Volume*, pages 341–383. Springer.

Albert, M., Bouret, Y., Fromont, M., and Reynaud-Bouret, P. (2015). Bootstrap and permutation tests of independence for point processes. *The Annals of Statistics*, 43(6):2537 – 2564.

Arias-Castro, E., Pelletier, B., and Saligrama, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471.

Ba, K. D., Nguyen, H. L., Nguyen, H. N., and Rubinfeld, R. (2011). Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, 48(2):428–442.

Balakrishnan, N. and Nevzorov, V. B. (2004). *A primer on statistical distributions.* John Wiley & Sons.

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577 – 606.

Biggs, F., Schrab, A., and Gretton, A. (2023). Mmd-fuse: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36:75151–75188.

Bobkov, S. G. and Ledoux, M. (2021). A simple fourier analytic proof of the akt optimal matching theorem. *The Annals of Applied Probability*, 31(6):2567–2584.

Boedihardjo, M. T. (2025). Sharp bounds for max-sliced wasserstein distances. *Foundations of Computational Mathematics*, pages 1–32.

Bonnotte, N. (2013). *Unidimensional and evolution methods for optimal transportation.* PhD thesis, ENS Cachan.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57.

Canonne, C. L. (2022). A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198.*

Chakravarti, P., Kuusela, M., Lei, J., and Wasserman, L. (2023). Model-independent detection of new physics signals using interpretable semisupervised classifier tests. *The Annals of Applied Statistics*, 17(4):2759–2795.

Chatalic, A., Letizia, M., Schreuder, N., and Rosasco, L. (2025). A scalable nyström-based kernel two-sample test with permutations. *arXiv preprint arXiv:2502.13570.*

Chewi, S., Niles-Weed, J., and Rigollet, P. (2025). *Statistical optimal transport.* Springer.

Choi, I. and Kim, I. (2024). Computational-statistical trade-off in kernel two-sample testing with random fourier features. *arXiv preprint arXiv:2407.08976.*

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE.

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

Domingo-Enrich, C., Dwivedi, R., and Mackey, L. (2023). Compress then test: Powerful kernel testing in near-linear time. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1174–1218. PMLR.

Dubhashi, D. P. and Ranjan, D. (1996). Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25).

Dümbgen, L. (1998). Symmetrization and decoupling of combinatorial random elements. *Statistics & probability letters*, 39(4):355–361.

El-Yaniv, R. and Pechyony, D. (2009). Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234.

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

Gao, I., Liang, P., and Guestrin, C. (2024). Model equality testing: Which model is this API serving? *arXiv preprint arXiv:2410.20247*.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

González-Delgado, J. A., Weed, J., and Del Barrio, E. (2023). Two-sample tests for distributions supported on the flat torus. *Annals of Statistics*.

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

Grossi, S., Letizia, M., and Torre, R. (2025). Refereeing the referees: evaluating two-sample tests for validating generators in precision sciences. *Machine Learning: Science and Technology*, 6(1):015052.

Hemerik, J. and Goeman, J. J. (2018). Exact testing with random permutations. *TEST*, 27(4):811–825.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23(2):169–192.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

Horváth, L., Reeder, R., and Kokoszka, P. (2014). Estimation of the mean of functional time series and a two-sample problem. *Quality control and applied statistics*, 59(1):51–52.

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

Hu, X. and Lin, Z. (2025). Two-sample distribution tests in high dimensions via max-sliced wasserstein distance and bootstrapping. *Biometrika*, 112(2):asaf001.

Imaizumi, M., Ota, H., and Hamaguchi, K. (2022). Hypothesis testing for distributions supported on general metric spaces with wasserstein distances. *arXiv preprint arXiv:2201.10456*.

Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist*, 2(2):85–114.

Jameson, G. (2015). A simple proof of stirling's formula for the gamma function. *The Mathematical Gazette*, 99(544):68–74.

Kim, I., Balakrishnan, S., and Wasserman, L. (2022). Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251.

Kim, I. and Schrab, A. (2023). Differentially private permutation tests: Applications to kernel methods. *arXiv preprint arXiv:2310.19043*.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kolouri, S., Rohde, G. K., and Hoffmann, H. (2019). Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems*.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

LeCun, Y., Cortes, C., and Burges, C. J. C. (2012). The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`. [Online; accessed 25-April-2018].

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30.

Liu, Y., Li, C.-L., and Póczos, B. (2018). Classifier two sample test for video anomaly detections. In *BMVC*, page 71.

Lopes, M. E., Jacob, L., and Wainwright, M. J. (2011). More powerful two-sample tests via random projections. *Neural Information Processing Systems*.

McDiarmid, C. et al. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

Miles, R. R., Roberts, R. F., Putnam, A. R., and Roberts, W. L. (2004). Comparison of serum and heparinized plasma samples for measurement of chemistry analytes. *Clinical chemistry*, 50(9):1704–1706.

Mukherjee, S. and Sriperumbudur, B. K. (2025). Minimax optimal kernel two-sample tests with random features. *arXiv preprint arXiv:2502.20755*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Nadjahi, K. (2021). *Sliced-Wasserstein distance for large-scale machine learning: theory, methodology and extensions*. PhD thesis, Institut polytechnique de Paris.

Nguyen, K. (2025). An introduction to sliced optimal transport: foundations, advances, extensions, and applications. *Foundations and Trends in Computer Graphics and Vision*, 17(3-4):171–391.

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

Ohana, R., Nadjahi, K., Rakotomamonjy, A., and Ralaivola, L. (2023). Shedding a pac-bayesian light on adaptive sliced-wasserstein distances. In *International Conference on Machine Learning*, pages 26451–26473. PMLR.

Paty, F.-P. and Cuturi, M. (2019). Subspace robust wasserstein distances. In *International Conference on Machine Learning*, pages 5072–5081.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). Subsampling in the iid case. In *Subsampling*, pages 39–64. Springer.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and sliced wasserstein distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer.

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

Rodríguez-Vítores, D., del Barrio, E., and Loubes, J.-M. (2025). An improved central limit theorem for the empirical sliced wasserstein distance. *arXiv preprint arXiv:2503.18831*.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). Mmd aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Shekhar, S., Kim, I., and Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, 35:18168–18180.

Shekhar, S., Kim, I., and Ramdas, A. (2023). A permutation-free kernel independence test. *Journal of Machine Learning Research*, 24(369):1–68.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2):279–281.

Soch, J. et al. (2025). Statproofbook/statproofbook.github.io: The book of statistical proofs (version 2024).

Stelzer, J., Chen, Y., and Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (mvpa): random permutations and cluster size control. *Neuroimage*, 65:69–82.

Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*.

Székely, G. J., Rizzo, M. L., et al. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.

Tolstikhin, I. (2017). Concentration inequalities for samples without replacement. *Theory of Probability & Its Applications*, 61(3):462–481.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pages 16–28. Springer.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Wang, J., Boedihardjo, M., and Xie, Y. (2025). Statistical and computational guarantees of kernel max-sliced Wasserstein distances. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J., editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 62373–62400. PMLR.

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

Wang, J., Gao, R., and Xie, Y. (2022). Two-sample test with kernel projected wasserstein distance. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8022–8055. PMLR.

Wasserman, L. (2014). Minimax theory. Lecture notes, Carnegie Mellon University. Section 11: Minimax Lower Bounds.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

Wei, F. and Dudley, R. M. (2012). Two-sample dvoretzky–kiefer–wolfowitz inequalities. *Statistics & Probability Letters*, 82(3):636–644.

Wendel, J. G. (1948). Note on the gamma function. *The American Mathematical Monthly*, 55(9):563.

Xu, X. and Huang, Z. (2022). Central limit theorem for the sliced 1-wasserstein distance and the max-sliced 1-wasserstein distance. *arXiv preprint arXiv:2205.14624*.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in neural information processing systems*, 26.

Zhang, J., Merikangas, K. R., Li, H., and Shou, H. (2021). Two-sample tests for repeated measurements of histogram objects with applications to wearable device data. *arXiv preprint arXiv:2106.12768*.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Material
# Minimax-Optimal Two-Sample Test with Sliced Wasserstein

## OVERVIEW

This supplementary material contains additional discussions, detailed proofs, and an extra experiment that were omitted from the main text due to page constraints. The remaining sections are organized as follows.

- In Section 8, we recall several key concentration inequalities that serve as important tools for our proofs.
- In Section 9, we derive intermediate technical lemmas for permuted sliced Wasserstein statistics using results from optimal transport theory. A core contribution of this section is the derivation of a novel concentration inequality and an expectation bound for the permuted sliced Wasserstein statistic. These results are fundamental for controlling the random permutation threshold in our power analysis.
- In Section 10, we present a result characterizing the sample complexity of the sliced Wasserstein distance.
- Our main theoretical contributions are presented in Sections 11 and 12. These sections provide the complete proofs of our main theorems concerning the test's performance (Theorems 4 and 5) and its minimax optimality (Propositions 6 and 7).
- In Section 13, we elaborate on why we chose a permutation-based framework for two-sample testing with the sliced Wasserstein distance, highlighting the intractability and practical challenges that prevent the use of the statistic's asymptotic null distribution.
- Section 14 offers a technical review of existing methods for analyzing permutation tests. We explain why common techniques developed for U-statistics and other test statistics are not directly applicable to the sliced Wasserstein distance, thereby underscoring the novelty of our analytical approach.
- Finally, Section 15 provides results for the Gaussian mean shift experiment mentioned in Section 6.

**Additional Notation.** Throughout this supplementary material, we use an additional set of notation, described below.

- For $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer greater than or equal to $x$, and $\lfloor x \rfloor$ denotes the largest integer less than or equal to $x$.
- We write $d_{\mathrm{TV}}(P, Q)$ for the total variation (TV) distance between $P$ and $Q$, and $D_{\mathrm{KL}}(P\|Q)$ for the Kullback–Leibler (KL) divergence. More details on those metrics can be found in, e.g., Tsybakov (2008, Section 2.4).
- For any integer $K \geq 1$, we denote by $[K]$ the set of the first $K$ integers $\{1, \ldots, K\}$.
- We use $\mathbb{P}_{\mu\times\nu}$ to denote probability with respect to $Y_1, \ldots, Y_n \sim \mu$ and $Z_1, \ldots, Z_m \sim \nu$. When additional randomness is present, we extend the notation accordingly, e.g., $\mathbb{P}_{\mu\times\nu\times r\times\sigma}$.
- Finally, given two random variables $U$ and $V$, we write $U \overset{(d)}{=} V$ to denote that $U$ and $V$ have the same distribution.

# 8 SOME USEFUL CONCENTRATION INEQUALITIES

We begin this section by recalling two classical concentration inequalities: McDiarmid's inequality (McDiarmid et al., 1989) and Hoeffding's inequality (Hoeffding, 1963).

**Lemma 8 (McDiarmid's Inequality)** *Let $N \geq 1$. Let $X_1, \ldots, X_N$ be independent random variables taking values in a set $\mathcal{X}$, and let $f : \mathcal{X}^N \to \mathbb{R}$. Assume that $f$ satisfies the* bounded difference property*, i.e., there exist constants $c_1, \ldots, c_N \geq 0$ such that for all $i \in \{1, \ldots, N\}$ and for all $x_1, \ldots, x_N, x_i' \in \mathcal{X}$,*

$$\big| f(x_1, \ldots, x_i, \ldots, x_N) - f(x_1, \ldots, x_i', \ldots, x_N) \big| \leq c_i.$$

*Then, for any $t > 0$,*

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

**Lemma 9 (Hoeffding's Inequality)** *Let $N \geq 1$. Let $X_1, \ldots, X_N$ be independent random variables such that $a_i \leq X_i \leq b_i$ for $i = 1, \ldots, N$. Then, for any $t > 0$,*

$$\mathbb{P}\left[ \left| \frac{1}{N} \sum_{i=1}^N X_i - \mathbb{E}\left( \frac{1}{N} \sum_{i=1}^N X_i \right) \right| \geq t \right] \leq 2 \exp\left( -\frac{2N^2 t^2}{\sum_{i=1}^N (b_i - a_i)^2} \right).$$

We shall also rely on a permutation-based version of the classical McDiarmid's inequality to control the empirical quantile $\widehat{c}^B_{1-\alpha,N}$ introduced in Algorithm 1. A statement of this result can be found in Tolstikhin (2017, Theorem 6) and El-Yaniv and Pechyony (2009, Lemma 2). To prepare for the statement of that result, we first introduce the following definition.

**Definition 10 ($(n,m)$-symmetric function)** *Let $N \geq 1$ and let $n$ be a positive integer such that $n < N$. Set $m := N - n$. A function $f : S_N \to \mathbb{R}$, defined on the symmetric group over $\{1, \ldots, N\}$, is called $(n,m)$-symmetric if it remains invariant under the change of order of the first $n$ coordinates and/or last $m$ coordinates of any permutation $\pi \in S_N$.*

We are now ready to state a permutation-based version of McDiarmid's inequality.

**Lemma 11 (Permutation McDiarmid's inequality)** *Let $n$ and $N$ be positive integers with $n < N$. Let $\pi$ be uniformly distributed on the symmetric group of $\{1, \ldots, N\}$, and let $f$ be an $(n, N-n)$-symmetric function. Assume that there exists a constant $b > 0$ such that*

$$|f(\pi) - f(\pi^{i,j})| \leq b, \qquad \forall \pi,\ \forall i \in \{1, \ldots, n\},\ \forall j \in \{n+1, \ldots, N\},$$

*where $\pi^{i,j}$ is obtained from $\pi$ by swapping its $i$-th and $j$-th entries. Then, for any $\epsilon > 0$,*

$$\mathbb{P}_\pi \left( f(\pi) - \mathbb{E}_\pi[f(\pi)] \geq \epsilon \right) \leq \exp\left( -\frac{2\epsilon^2}{nb^2} \cdot \frac{N - \frac{1}{2}}{N - n} \cdot \left( 1 - \frac{1}{2\max\{n, N-n\}} \right) \right).$$

# 9 PROPERTIES OF THE PERMUTED SLICED WASSERSTEIN STATISTIC

Our main objective in this section is to obtain a deterministic upper bound the (random) permuted test statistics $\widehat{\mathrm{SW}}_p^{p,\pi}$ introduced in (4). To this end, we exploit the properties of the sliced Wasserstein distance under permutation via a permutation-based version of McDiarmid's inequality (recalled in Lemma 11). It yields a high-probability bound relating the permuted statistics to its expectation. We then control the expectation term by applying the optimal matching bound from Bobkov and Ledoux (2021, Corollary 5).

We recall that all results are established under the assumption that the samples $\mathcal{Y}_n = (y_1, \ldots, y_n)$ and $\mathcal{Z}_m = (z_1, \ldots, z_m)$ satisfy $\|y_i\| \leq D$ and $\|z_j\| \leq D$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Moreover, $\Theta = (\theta_1, \ldots, \theta_L)$ denotes the projection directions drawn from the uniform distribution $\sigma$ on the unit sphere $\mathbb{S}^{d-1}$.

We first establish a high-probability concentration inequality for the permuted empirical sliced Wasserstein distance $\widehat{\mathrm{SW}}_p^{p,\pi}$ introduced in (4). It is the first step towards proving our main theorem in Section 11.2.

**Lemma 12** *Let $p \geq 1$. Let $\pi$ be a permutation drawn uniformly from the symmetric group on $\{1, \ldots, n+m\}$, and let $\pi^{i,j}$ denote the permutation obtained from $\pi$ by exchanging its $i$-th and $j$-th elements, as defined in Lemma 11. Then, for any $\epsilon > 0$,*

$$\mathbb{P}_\pi\left(\widehat{\mathrm{SW}}_p^{p,\pi} - \mathbb{E}_\pi\left[\widehat{\mathrm{SW}}_p^{p,\pi}\right] \geq \epsilon \,\middle|\, \mathcal{Y}_n, \mathcal{Z}_m, \Theta\right) \leq \exp\left(-\frac{2\epsilon^2}{n\left(\frac{(2D)^p}{n} + \frac{(2D)^p}{m}\right)^2} \cdot \frac{n+m-\frac{1}{2}}{m}\left(1 - \frac{1}{2\max\{m,n\}}\right)\right).$$

**Proof** Given the samples $\mathcal{Y}_n, \mathcal{Z}_m$ and a set of projection directions $\Theta = (\theta_1, \ldots, \theta_L)$ on the unit sphere $\mathbb{S}^{d-1}$, for brevity, we set

$$f(\pi) \coloneqq \widehat{\mathrm{SW}}_p^{p,\pi}, \qquad \pi \in S_{n+m},$$

where $\widehat{\mathrm{SW}}_p^{p,\pi}$ is the permuted empirical quantity introduced in (4).

By construction, $f$ is $(n,m)$-symmetric. In order to apply Lemma 11 we need to get an upper bound on the sensitivity of $f$ to transpositions involving elements from opposite halves of the permutation.

Without loss of generality, suppose that $\pi = (1, 2, 3, \ldots, n+m)$ and $\pi^{i,j} = (1, 2, \ldots, i-1, j, i+1, \ldots, n, \ldots, j-1, i, j+1, \ldots n+m)$ for $1 \leq i \leq n$ and $n+1 \leq j \leq n+m$.

Let $\ell \in \{1, \ldots, L\}$. Denote the projected samples on the direction $\theta_\ell$ as

$$\tilde{\mathcal{X}} \coloneqq (\tilde{x}_1, \ldots, \tilde{x}_n, \tilde{x}_{n+1}, \ldots, \tilde{x}_{n+m}) = (\langle y_1, \theta_\ell\rangle, \ldots, \langle y_n, \theta_\ell\rangle, \langle z_1, \theta_\ell\rangle, \ldots, \langle z_m, \theta_\ell\rangle).$$

By definition of the $p$-Wasserstein distance and the permutation $\pi$, we have

$$\begin{aligned}
\mathrm{W}_p^p\left(\Pi_\#^{\theta_\ell}\widehat{\mu}_n^\pi, \Pi_\#^{\theta_\ell}\widehat{\nu}_m^\pi\right) &= \min_{\substack{\gamma \geq 0, \\ \sum_k \gamma_{kl} = \frac{1}{m}, \\ \sum_l \gamma_{kl} = \frac{1}{n}}} \left(\sum_{k=1}^n \sum_{l=1}^m \gamma_{kl} \left|\tilde{x}_{\pi(k)} - \tilde{x}_{\pi(n+l)}\right|^p\right) \\
&= \min_{\substack{\gamma \geq 0, \\ \sum_k \gamma_{kl} = \frac{1}{m}, \\ \sum_l \gamma_{kl} = \frac{1}{n}}} \left(\sum_{k=1}^n \sum_{l=1}^m \gamma_{kl} \left|\tilde{x}_k - \tilde{x}_{n+l}\right|^p\right) \\
&= \sum_{k,l} \gamma_{kl}^* \left|\tilde{x}_k - \tilde{x}_{n+l}\right|^p,
\end{aligned}$$

where $\gamma^*$ denotes the minimizer of the above Kantorovich problem.

We now consider the permutation $\pi^{i,j}$. Noting that $\gamma^*$ is a valid coupling, we have

$$\begin{aligned}
\mathrm{W}_p^p\left(\Pi_\#^{\theta_\ell}\widehat{\mu}_n^{\pi^{i,j}}, \Pi_\#^{\theta_\ell}\widehat{\nu}_m^{\pi^{i,j}}\right) &= \min_{\substack{\gamma \geq 0, \\ \sum_k \gamma_{kl} = \frac{1}{m}, \\ \sum_l \gamma_{kl} = \frac{1}{n}}} \sum_{k=1}^n \sum_{l=1}^m \gamma_{kl} \left|\tilde{x}_{\pi^{i,j}(k)} - \tilde{x}_{\pi^{i,j}(n+l)}\right|^p \\
&\leq \sum_{k=1}^n \sum_{l=1}^m \gamma_{kl}^* \left|\tilde{x}_{\pi^{i,j}(k)} - \tilde{x}_{\pi^{i,j}(n+l)}\right|^p \\
&\leq \sum_{k=1}^n \sum_{l=1}^m \gamma_{kl}^* \left|\tilde{x}_k - \tilde{x}_{n+l}\right|^p \\
&\quad + \sum_{l=1}^m \gamma_{il}^* \left|\tilde{x}_{\pi^{i,j}(i)} - \tilde{x}_{\pi^{i,j}(n+l)}\right|^p \\
&\quad + \sum_{k=1}^n \gamma_{k,\,j-n}^* \left|\tilde{x}_{\pi^{i,j}(k)} - \tilde{x}_{\pi^{i,j}(j)}\right|^p \\
&\leq \mathrm{W}_p^p\left(\Pi_\#^{\theta_\ell}\widehat{\mu}_n^\pi, \Pi_\#^{\theta_\ell}\widehat{\nu}_m^\pi\right) + \frac{(2D)^p}{n} + \frac{(2D)^p}{m}.
\end{aligned}$$

where the last inequality follows from the triangle inequality and the fact that every projected sample satisfies

$$\left|\tilde{x}_h\right| = \left|\langle x_h, \theta_\ell\rangle\right| \le \|x_h\| \, \|\theta_\ell\| \le D, \quad \text{for } 1 \le h \le n+m,$$

with $(x_1, \ldots, x_{n+m})$ denoting the pooled sample.

As a result, we obtain

$$f(\pi^{i,j}) - f(\pi) \le \frac{(2D)^p}{n} + \frac{(2D)^p}{m}.$$

Following the same steps, we can prove that

$$f(\pi) - f(\pi^{i,j}) \le \frac{(2D)^p}{n} + \frac{(2D)^p}{m},$$

which is sufficient to conclude that

$$\left|f(\pi) - f(\pi^{i,j})\right| \le \frac{(2D)^p}{n} + \frac{(2D)^p}{m}.$$

Applying Lemma 11 with $b = \frac{(2D)^p}{n} + \frac{(2D)^p}{m}$ and $f(\pi) = \widehat{\mathrm{SW}}_p^{p,\pi}$ completes the proof. ■

Having successfully related the permuted test statistic $\widehat{\mathrm{SW}}_p^{p,\pi}$ (introduced in (4)) to its expectation, we now turn to controlling this expectation term. To proceed, we next present several auxiliary lemmas that will be used later in the proof.

Our argument begins with the following result, stated by Bobkov and Ledoux (2021, Corollary 5).

**Lemma 13 (Corollary 5 in Bobkov and Ledoux (2021))** *Let $n$ and $N$ be integers such that $1 \le n \le N$. Let $x_1, \ldots, x_N \in [0,1]^d$. Denote by $\mathcal{G}_n$ the collection of all subsets $\tau \subset \{1, \ldots, N\}$ of cardinality $|\tau| = n$, equipped with the uniform probability measure $\pi_n$. With every $\tau \in \mathcal{G}_n$, we associate the empirical measure*

$$\mu_\tau \coloneqq \frac{1}{n}\sum_{j\in\tau} \delta_{x_j}, \quad \tilde{\mu} \coloneqq \frac{1}{N}\sum_{j=1}^{N} \delta_{x_j}.$$

*Then the empirical measures $\mu_\tau$ satisfy*

$$\mathbb{E}_{\pi_n}\left[W_1(\mu_\tau, \tilde{\mu})\right] \le \begin{cases} \sqrt{\dfrac{2}{n}} & \text{if } d = 1, \\ 8\sqrt{\dfrac{1+\log(2n)}{n}} & \text{if } d = 2, \\ \dfrac{13\sqrt{d}}{n^{1/d}} & \text{if } d \ge 3. \end{cases}$$

**Remark 14** *Corollary 5 in Bobkov and Ledoux (2021) is stated for empirical measures supported on the unit cube $[0,1]^d$. We next extend this result to empirical measures supported on a general cube $[-D, D]^d$, where $D > 0$. To connect these two settings, we introduce the affine transformation*

$$T\colon [0,1]^d \to [-D, D]^d, \qquad T(x) = 2D(x - \mathbf{1}_d),$$

*where $\mathbf{1}_d \coloneqq (1, 1, \ldots, 1) \in \mathbb{R}^d$. Given data points $x_1, \ldots, x_N \in [0,1]^d$, we define*

$$x'_i = T(x_i), \quad i = 1, \ldots, N,$$

*so that each $x'_i$ lies in the cube $[-D, D]^d$.*

*Let $\mu'_\tau$ and $\tilde{\mu}'$ denote the empirical measures defined analogously to $\mu_\tau$ and $\tilde{\mu}$ in Lemma 13 but based on points $(x'_i)_{i=1}^N$, that is*

$$\mu'_\tau = \frac{1}{n}\sum_{j\in\tau} \delta_{x'_j}, \qquad \tilde{\mu}' = \frac{1}{N}\sum_{j=1}^{N} \delta_{x'_j}.$$

*By construction, we have $T_{\#}\mu_\tau = \mu'_\tau$ and $T_{\#}\tilde{\mu} = \tilde{\mu}'$. Moreover, for any $x, y \in [0,1]^d$, we have*

$$\|T(x) - T(y)\|_2 = 2D\|x - y\|_2.$$

*Consequently, we have the scaling relation*

$$W_1(\mu'_\tau, \tilde{\mu}') = 2D\, W_1(\mu_\tau, \tilde{\mu}).$$

*Applying Corollary 5 in Bobkov and Ledoux (2021) to $\mu_\tau$ and $\tilde{\mu}$ then yields*

$$\mathbb{E}_{\pi_n}[W_1(\mu'_\tau, \tilde{\mu}')] = \mathbb{E}_{\pi_n}\left[(2D)W_1(\mu_\tau, \tilde{\mu})\right] \leq \begin{cases} 2D\sqrt{\dfrac{2}{n}} & \text{if } d = 1, \\ 16D\sqrt{\dfrac{1+\log(2n)}{n}} & \text{if } d = 2, \\ \dfrac{26D\sqrt{d}}{n^{1/d}} & \text{if } d \geq 3. \end{cases}$$

**Lemma 15 (Relation between 1-Wasserstein and $p$-Wasserstein distances)** *Let $P$, $Q$, and $M$ be three probability measures supported on the compact set $[-D, D]^d \subset \mathbb{R}^d$. Let $p \geq 1$, then the following inequality holds:*

$$\mathrm{W}_p^p(P, Q) \leq (4D\sqrt{d})^{p-1}\left(\mathrm{W}_1(P, Q) + \mathrm{W}_1(Q, M)\right).$$

**Proof** By applying the triangle inequality for the $p$-Wasserstein distance, we get

$$\mathrm{W}_p(P, Q) \leq \mathrm{W}_p(P, Q) + \mathrm{W}_p(Q, M).$$

Using the inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p)$, we obtain

$$\mathrm{W}_p^p(P, Q) \leq 2^{p-1}\left(\mathrm{W}_p^p(P, M) + \mathrm{W}_p^p(Q, M)\right).$$

For $x, y \in [-D, D]^d$, we have

$$\|x - y\|_2 \leq 2D\sqrt{d}.$$

Moreover, for $p \geq 1$, we have

$$\|x - y\|_2^p \leq (2D\sqrt{d})^{p-1}\|x - y\|_2.$$

Hence, since all three measures are supported on $[-D, D]^d$, by the definition of Wasserstein distances, we obtain

$$\mathrm{W}_p^p(P, M) \leq (2D\sqrt{d})^{p-1}\,\mathrm{W}_1(P, M), \quad \mathrm{W}_p^p(Q, M) \leq (2D\sqrt{d})^{p-1}\,\mathrm{W}_1(Q, M).$$

Combining all the inequalities gives the stated result. ■

**Remark 16** *We recall that our standing assumption is that all samples $\mathcal{Y}_n = (y_1, \ldots, y_n)$ and $\mathcal{Z}_m = (z_1, \ldots, z_m)$ satisfy $\|y_i\| \leq D$ and $\|z_j\| \leq D$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Meanwhile, the results in Lemma 13 and Lemma 15 are stated for the case where all data points lie in the cube $[-D, D]^d$. Since the Euclidean ball $\{x \in \mathbb{R}^d : \|x\| \leq D\}$ is contained in this cube, those results remain valid in our setting.*

Using the results of Lemma 13 and Lemma 15, along with their accompanying remarks, the following proposition establishes an upper bound for the expectation term conditional on the given samples and the projection directions.

**Proposition 17** *Let $n$ and $m$ be positive integers such that $n \leq m$, $p \geq 1$ and set $N := n+m$. Given a collection of points $x_1, \ldots, x_N$ with $\|x_i\| \leq D$ for all $1 \leq i \leq N$, and a set of projection directions $\Theta = (\theta_1, \ldots, \theta_L)$ on the unit sphere $\mathbb{S}^{d-1}$. Let $\pi$ be a random permutation drawn uniformly from the symmetric group over $\{1, \ldots, N\}$. Then, with $p \geq 1$, we have*

$$\mathbb{E}_\pi\left[\widehat{\mathrm{SW}}_p^{p,\pi}\right] = \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_\pi\left[\mathrm{W}_p^p\Big(\Pi_{\#}^{\theta_\ell}\widehat{\mu}_n^\pi,\, \Pi_{\#}^{\theta_\ell}\widehat{\nu}_m^\pi\Big)\right] \leq (4D)^p\frac{\sqrt{2}}{\sqrt{n}}.$$

*Here, the expectation $\mathbb{E}_\pi$ is taken with respect to the random permutation $\pi$, conditional on the samples and the projection directions.*

**Proof** Fix a projection direction $\theta_\ell$, $1 \le \ell \le L$. We want to upper bound $\mathbb{E}_\pi\left[\mathrm{W}_p^p\Big(\Pi_\#^{\theta_\ell}\widehat{\mu}_n^\pi,\, \Pi_\#^{\theta_\ell}\widehat{\nu}_m^\pi\Big)\right]$.

We introduce the empirical measure

$$\tilde{\mu} := \frac{1}{N}\sum_{j=1}^{N}\delta_{\langle x_j,\theta_\ell\rangle}.$$

Note that each projected sample satisfies $|\langle x_j,\theta_\ell\rangle| \le D$ for all $1 \le j \le N$ by the Cauchy–Schwarz inequality.

Adapting the notations from Lemma 13, we first define $\mathcal{G}_n$ and $\mathcal{G}_m$ as the collections of all subsets of $\{1,\dots,n{+}m\}$ of cardinalities $n$ and $m$, respectively. We denote by $\pi_n$ and $\pi_m$ the uniform distributions on $\mathcal{G}_n$ and $\mathcal{G}_m$. To every $\tau^n \in \mathcal{G}_n$ and $\tau^m \in \mathcal{G}_m$, we associate two empirical measures

$$\mu_{\tau^n} := \frac{1}{n}\sum_{j\in\tau^n}\delta_{\langle\theta_\ell,x_j\rangle}, \qquad \mu_{\tau^m} := \frac{1}{m}\sum_{j\in\tau^m}\delta_{\langle\theta_\ell,x_j\rangle}.$$

It follows that

$$\begin{aligned}
\mathbb{E}_\pi\left[\mathrm{W}_p^p\Big(\Pi_\#^{\theta_\ell}\widehat{\mu}_n^\pi,\, \Pi_\#^{\theta_\ell}\widehat{\nu}_m^\pi\Big)\right] &= \frac{1}{(n+m)!}\sum_{\pi\in S_{n+m}}\mathrm{W}_p^p\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\langle x_{\pi(i)},\theta_\ell\rangle},\frac{1}{m}\sum_{i=1}^{m}\delta_{\langle x_{\pi(n+i)},\theta_\ell\rangle}\right)\\
&= \frac{m!.n!}{(m+n)!}\sum_{K\subset\{x_1,\dots,x_{n+m}\},|K|=n}\mathrm{W}_p^p\left(\frac{1}{n}\sum_{x\in K}\delta_{\langle x,\theta_\ell\rangle},\frac{1}{m}\sum_{x\in K^c}\delta_{\langle x,\theta_\ell\rangle}\right)\\
&\overset{(i)}{\le} \frac{(4D)^{p-1}}{\binom{n+m}{n}}\sum_{K\subset\{x_1,\dots,x_{n+m}\},|K|=n}\left[\mathrm{W}_1\left(\frac{1}{n}\sum_{x\in K}\delta_{\langle x,\theta_\ell\rangle},\tilde{\mu}\right)+\mathrm{W}_1\left(\frac{1}{m}\sum_{x\in K^c}\delta_{\langle x,\theta_\ell\rangle},\tilde{\mu}\right)\right]\\
&= (4D)^{p-1}\left[\mathbb{E}_{\pi_m}(\mathrm{W}_1(\mu_{\tau^m},\tilde{\mu}))+\mathbb{E}_{\pi_n}(\mathrm{W}_1(\mu_{\tau^n},\tilde{\mu}))\right]\\
&\overset{(ii)}{\le} 2D\cdot(4D)^{p-1}\left(\frac{\sqrt{2}}{\sqrt{n}}+\frac{\sqrt{2}}{\sqrt{m}}\right)\\
&\overset{(iii)}{\le} 4D\cdot(4D)^{p-1}\frac{\sqrt{2}}{\sqrt{n}},
\end{aligned}$$

where $(i)$ follows from Lemma 15, and $(ii)$ and $(iii)$ follow from Lemma 13 and from the assumption that $n \le m$, respectively.

Finally, by linearity of the expectation,

$$\mathbb{E}_\pi\left[\widehat{\mathrm{SW}}_p^{p,\pi}\right] = \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_\pi\left[\mathrm{W}_p^p\big((\Pi^{\theta_\ell})_\#\widehat{\mu}_n^\pi,\,(\Pi^{\theta_\ell})_\#\widehat{\nu}_m^\pi\big)\right] \le (4D)^p\frac{\sqrt{2}}{\sqrt{n}},$$

and the proof is concluded. ■

## 10 SLICED WASSERSTEIN SAMPLE COMPLEXITY

Let $\mu$ and $\nu$ be probability distributions on $\mathbb{R}^d$ whose supports lie in a common ball centered at the origin with radius $D$. Their corresponding empirical measures are denoted by $\widehat{\mu}_n$ and $\widehat{\nu}_m$. The quantities $\mathrm{SW}_p^p(\mu,\nu)$ and $\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m)$ are defined in Eqs. (1) and (2), respectively.

To the best of our knowledge, recent works have established bounds on quantities such as $\mathbb{E}\left[\mathrm{SW}_p(\widehat{\mu}_n,\mu)\right]$ (see, e.g., Nietert et al. (2022, Theorem 1)), $\mathbb{E}\left[\mathrm{SW}_p^p(\widehat{\mu}_n,\widehat{\nu}_n)-\mathrm{SW}_p^p(\mu,\nu)\right]$ (see, e.g., Ohana et al. (2023, Theorem 2)), and $\mathbb{E}\left[\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m)-\mathrm{SW}_p^p(\mu,\nu)\right|\right]$ (see, e.g., Nietert et al. (2022, Proposition 5)), as well as high-probability bounds for $\left|\mathrm{SW}_p^p(\mu,\nu)-\widehat{\mathrm{SW}}_p^p(\mu,\nu)\right|$ (see, e.g., Xu and Huang (2022, Proposition 4)). These results, however, do not exactly yield the type of bound required for the subsequent step—specifically, in the proof of Lemma 19.

Under the assumption that both distributions $\mu$ and $\nu$ are supported on a common bounded ball, we derive the following high-probability bound for $\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_p^p(\mu, \nu)\right|$ by applying standard concentration inequalities—namely McDiarmid's inequality and Hoeffding's inequality (recalled in Lemma 8 and Lemma 9, respectively) and following the technique established in Nietert et al. (2022, Proposition 1). The resulting statement is presented in the following lemma.

**Lemma 18** *Let $\beta \in (0,1)$ and $p \geq 1$. Suppose that $\mu$ and $\nu$ are probability distributions supported on the centered ball of radius $D > 0$ in $\mathbb{R}^d$. We consider two independent samples drawn from $\mu$ and $\nu$:*

$$\mathcal{Y}_n \coloneqq (Y_1, \ldots, Y_n) \overset{i.i.d.}{\sim} \mu, \qquad \mathcal{Z}_m \coloneqq (Z_1, \ldots, Z_m) \overset{i.i.d.}{\sim} \nu,$$

*with an assumption that $n \leq m$. Let $\Theta \coloneqq (\theta_1, \ldots, \theta_L)$ denote $L$ i.i.d. projection directions drawn from the uniform distribution $\sigma$ on the unit sphere $\mathbb{S}^{d-1}$. Then, there exists a constant $C_{p,D}$ depending only on $p$ and $D$ such that, with probability at least $1 - \frac{\beta}{2}$, the following inequality holds:*

$$\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_p^p(\mu, \nu)\right| \;<\; (2D)^p \left(\sqrt{\frac{\log(8/\beta)}{2L}} + \sqrt{\frac{\log(8/\beta)}{n}}\right) + \frac{C_{p,D}}{\sqrt{n}}.$$

**Proof** We decompose the total error into three contributions:

$$\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_p^p(\mu, \nu)\right| \;\leq\; A + B + C,$$

where

$$A = \left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m)\right|,\; B = \left|\,\mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathbb{E}\,\mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m)\right|,\; C = \mathbb{E}\left|\mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_{\mathrm{p}}^{\mathrm{p}}(\mu, \nu)\right|.$$

**Control of $A$.** Given the datasets $\mathcal{Y}_n, \mathcal{Z}_m$, the Monte Carlo estimator $\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m)$ is an average of $L$ i.i.d. bounded random variables

$$R_\ell = \mathrm{W}_p^p\big(\Pi_\#^{\theta_\ell}\widehat{\mu}_n, \Pi_\#^{\theta_\ell}\widehat{\nu}_m\big), \qquad \ell = 1, \ldots, L.$$

Since $\|\theta_\ell\|_2 = 1$ and the supports of $\mu$ and $\nu$ are contained in the centered ball with radius $D$, the push-forward measures $(\Pi^{\theta_\ell})_\#\widehat{\mu}_n$ and $(\Pi^{\theta_\ell})_\#\widehat{\nu}_m$ are supported on the interval $[-D, D]$. Consequently,

$$0 \;\leq\; R_\ell \;\leq\; (2D)^p, \qquad \forall\, 1 \leq \ell \leq L.$$

Then, Hoeffding's inequality (recalled in Lemma 9) yields that, for any $t_A > 0$,

$$\mathbb{P}_\sigma(A \geq t_A \mid \mathcal{Y}_n, \mathcal{Z}_m) \;\leq\; 2\exp\!\left(-\frac{2Lt_A^2}{(2D)^{2p}}\right).$$

Since the right-hand side does not depend on the samples, the same bound holds unconditionally, namely

$$\mathbb{P}_{\mu,\nu,\sigma}(A \geq t_A) \;\leq\; 2\exp\!\left(-\frac{2Lt_A^2}{(2D)^{2p}}\right).$$

**Control of $B$.** Fix the projection directions $\theta_1, \ldots, \theta_L$. The function $(y_1, \ldots, y_n, z_1, \ldots, z_m) \mapsto \mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m)$ satisfies the bounded difference property, whose definition is recalled in Lemma 8. Indeed, changing a single observation $y_i$ modifies $\widehat{\mu}_n$ by at most $1/n$ of its mass. Since all points are supported in a set of diameter $2D$, the value of $\mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m)$ changes by at most $(2D)^p/n$. Similarly, replacing one sample $z_j$ alters $\widehat{\nu}_m$ by $1/m$ of its mass, and hence $\mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m)$ changes by at most $(2D)^p/m$. McDiarmid's inequality (recalled in Lemma 8) guarantees that, for any $t_B > 0$,

$$\mathbb{P}_{\mu,\nu}\left(B \geq t_B \mid \theta_1, \ldots, \theta_L\right) \leq 2\exp\!\left(-\frac{2t_B^2}{(2D)^{2p}(1/n + 1/m)}\right).$$

Since the right-hand side is independent of the projection directions, the bound also holds without conditioning, that is,

$$\mathbb{P}_{\mu,\nu,\sigma}\left(B \geq t_B\right) \;\leq\; 2\exp\left(-\frac{2t_B^2}{(2D)^{2p}(1/n+1/m)}\right).$$

Moreover, since $n \leq m$, then

$$\mathbb{P}_{\mu,\nu,\sigma}(B \geq t_B) \;\leq\; 2\exp\left(-\frac{nt_B^2}{(2D)^{2p}}\right).$$

**Control of C.**

Following the proof of Nietert et al. (2022, Proposition 1), and assuming that $n \leq m$, then there exists a constant $C_{p,D}$ depending on $p$ and $D$ such that

$$\mathbb{E}\left|\mathrm{SW}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_\mathrm{p}^\mathrm{p}(\mu,\nu)\right| \leq \frac{C_{p,D}}{\sqrt{n}}$$

We note that if $A < t_A$ and $B < t_B$, then $\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_p^p(\mu,\nu)\right| < t_A + t_B + \frac{C_{p,D}}{\sqrt{n}}$. Consequently, we obtain the following chain of evaluation

$$\begin{aligned}
\mathbb{P}_{\mu,\nu,\sigma}\left(\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SW}_p^p(\mu,\nu)\right| < t_A + t_B + \frac{C_{p,D}}{\sqrt{n}}\right) &\geq \mathbb{P}_{\mu,\nu,\sigma}\left(A < t_A, B < t_B\right) \\
&\geq 1 - \mathbb{P}_{\mu,\nu,\sigma}(A \geq t_A) - \mathbb{P}_{\mu,\nu,\sigma}(B \geq t_B) \\
&\geq 1 - 2\exp\left(-\frac{2Lt_A^2}{(2D)^{2p}}\right) - 2\exp\left(-\frac{nt_B^2}{(2D)^{2p}}\right)
\end{aligned}$$

We obtain the stated result by setting

$$t_A \;=\; (2D)^p\sqrt{\frac{\log(8/\beta)}{2L}} \quad \text{and} \quad t_B \;=\; (2D)^p\sqrt{\frac{\log(8/\beta)}{n}}.$$

■

# 11 LEVEL AND POWER GUARANTEES FOR ALGORITHM 1

## 11.1 Proof of Theorem 4

Following the proof of Schrab et al. (2023, Proposition 1), we obtain the following chain of implications

$$\begin{aligned}
\Delta\left(\mathcal{Y}_n, \mathcal{Z}_m, \mathbb{Z}_B, \Theta\right) = 1 &\Rightarrow \widehat{\mathrm{SW}}_p^p > \widehat{c}_{1-\alpha,N}^B \\
&\Rightarrow \widehat{\mathrm{SW}}_p^p > \widehat{\mathrm{SW}}_p^{p,r_{\bullet\lceil(B+1)(1-\alpha)\rceil}} \\
&\Rightarrow \sum_{b=1}^{B+1} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} < \widehat{SW}_p^{p,\pi_{B+1}}\right) \geq \lceil(B+1)(1-\alpha)\rceil \\
&\Rightarrow B+1-\sum_{b=1}^{B+1} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} < \widehat{\mathrm{SW}}_p^{p,\pi_{B+1}}\right) \leq B+1-\lceil(B+1)(1-\alpha)\rceil \\
&\Rightarrow \sum_{b=1}^{B+1} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} \geq \widehat{\mathrm{SW}}_p^{p,\pi_{B+1}}\right) \leq \lfloor\alpha(B+1)\rfloor
\end{aligned}$$

$$\Rightarrow \sum_{b=1}^{B+1} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} \geq \widehat{SW}_p^{p,\pi_{B+1}}\right) \leq \alpha(B+1)$$

$$\Rightarrow \frac{1}{B+1}\left[1+\sum_{b=1}^{B} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} \geq \widehat{\mathrm{SW}}_p^{p,\pi_{B+1}}\right)\right] \leq \alpha,$$

where the fifth deduction follows from the fact that $B+1-\lceil(1-\alpha)(B+1)\rceil = \lfloor\alpha(B+1)\rfloor$. Besides, the notations $\widehat{\mathrm{SW}}_p^{p}$ and $\widehat{\mathrm{SW}}_p^{p,\pi}$ refer respectively to the statistics defined in (2) and (4).

Moreover, Lemma 1 in Romano and Wolf (2005) guarantees that:

$$\mathbb{P}_{\mu\times\mu\times r\times\sigma}\left\{\frac{1}{B+1}\left[1+\sum_{b=1}^{B} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} \geq \widehat{\mathrm{SW}}_p^{p,\pi_{B+1}}\right)\right] \leq \alpha\right\} \leq \alpha.$$

As a consequence, we have

$$\mathbb{P}_{\mu\times\mu\times r\times\sigma}\left(\Delta\left(\mathcal{Y}_n, \mathcal{Z}_m, \mathbb{Z}_B, \Theta\right)=1\right) \leq \mathbb{P}_{\mu\times\mu\times r\times\sigma}\left\{\frac{1}{B+1}\left[1+\sum_{b=1}^{B} \mathbf{1}\left(\widehat{\mathrm{SW}}_p^{p,\pi_b} \geq \widehat{\mathrm{SW}}_p^{p,\pi_{B+1}}\right)\right] \leq \alpha\right\} \leq \alpha.$$

Hence, the test controls the Type I error at level $\alpha$.

### 11.2 Proof of Theorem 5

Let $r$ denote the uniform distribution over the symmetric group $S_{n+m}$, i.e., the set of all permutations of $\{1,\ldots,n+m\}$. Let $\pi$ be a permutation drawn from $r$. Moreover, we defined $c_{1-\alpha,N}$ as the $(1-\alpha)$-quantile (with respect to the randomness of $\pi$) of the test statistics $\widehat{\mathrm{SW}}_p^{p,\pi}$ (introduced in Eq. (4)), that is,

$$c_{1-\alpha,N} = \inf\left\{t : \mathbb{P}_\pi\left(\widehat{\mathrm{SW}}_p^{p,\pi} \geq t \,\middle|\, \mathcal{Y}_n, \mathcal{Z}_m, \Theta\right) \leq \alpha\right\}. \tag{6}$$

As explained in Remark 2, in practice we draw $B$ independent permutations $(\pi_b)_{1\leq b\leq B}$ from the uniform distribution $r$ on $S_{n+m}$ and set $\pi_{B+1} := \mathrm{id}$ to denote the identity permutation. We then compute the corresponding statistics $\widehat{\mathrm{SW}}_p^{p,\pi_b}$ for $b=1,\ldots,B+1$ and estimate the $(1-\alpha)$ empiraical quantile of the permutation distribution by

$$\widehat{c}^{B}_{1-\alpha,N} := \inf\left\{t : \frac{1}{B+1}\sum_{i=1}^{B+1} \mathbf{1}\{\widehat{\mathrm{SW}}_p^{p,\pi_i} \leq t\} \geq 1-\alpha\right\}.$$

In accordance with the convention introduced earlier, we use the notation $\mathbb{P}_r$ to denote the probability with respect to the $B$ random permutations $\pi_1,\ldots,\pi_B \overset{\text{i.i.d.}}{\sim} r$.

With these preliminaries in place, the proof of Theorem 5 proceeds as follows. Taking inspiration from Schrab et al. (2023, Lemma 4), we first provide a condition on the separation between the distributions $\mu$ and $\nu$ that guarantees a desired level of test power (see Lemma 19). This condition relates the separation to the (random) empirical quantile $\widehat{c}^{B}_{1-\alpha,N}$. Next, we control $\widehat{c}^{B}_{1-\alpha,N}$ by linking it to the (deterministic) quantile $c_{1-\alpha,N}$, conditional on the samples and the projection directions (see Lemma 20). Then, the quantile $c_{1-\alpha,N}$ is bounded using the auxiliary results prepared in Section 9 (see Proposition 21). Finally, we derive the bound for $\widehat{c}^{B}_{1-\alpha,N}$ stated in Proposition 22. We now develop this program in detail.

#### Main ingredients for the proof

We begin with the following lemma, which is a straightforward adaptation of Schrab et al. (2023, Lemma 4). It provides a sufficient condition on the separation between distributions $\mu$ and $\nu$ to ensure a desired level of test power.

**Lemma 19** *Let $\beta \in (0,1)$ and $p \geq 1$. Consider the setting of Lemma 18. Let*

$$\gamma(n,p,\beta,D,L) \coloneqq (2D)^p \left( \sqrt{\frac{\log(8/\beta)}{2L}} + \sqrt{\frac{\log(8/\beta)}{n}} \right) + \frac{C_{p,D}}{\sqrt{n}},$$

*where $C_{p,D}$ is a constant depending only on $p$ and $D$, as introduced in Lemma 18.*

*The test defined in Algorithm 1 achieves power at least $1-\beta$ provided that*

$$\mathbb{P}_{\mu\times\nu\times r\times\sigma}\left(\mathrm{SW}_p^p(\mu,\nu) \geq \gamma(n,p,\beta,D,L) + \widehat{c}^B_{1-\alpha,N}\right) > 1 - \tfrac{\beta}{2}. \tag{7}$$

**Proof** Lemma 18 guarantees

$$\mathbb{P}_{\mu\times\nu\times r\times\sigma}\Big(\big|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m) - \mathrm{SW}_p^p(\mu,\nu)\big| \;\geq\; \gamma(n,p,\beta,D,L)\Big) \;\leq\; \frac{\beta}{2}, \tag{8}$$

where

$$\gamma(n,p,\beta,D,L) \coloneqq (2D)^p \left( \sqrt{\frac{\log(8/\beta)}{2L}} + \sqrt{\frac{\log(8/\beta)}{n}} \right) + \frac{C_{p,D}}{\sqrt{n}}.$$

Define the events

$$\mathcal{A} \coloneqq \{\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m) \leq \widehat{c}^B_{1-\alpha,N}\} \quad \text{ and } \mathcal{B} \coloneqq \{\mathrm{SW}_p^p(\mu,\nu) \geq \gamma(n,p,\beta,D,L) + \widehat{c}^B_{1-\alpha,N}\}.$$

Let us show that $\mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{A}) \leq \beta$. By definition of the events and Eq. (8), we have

$$\begin{aligned}
\mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{A}\cap\mathcal{B}) &= \mathbb{P}_{\mu\times\nu\times r\times\sigma}\left(\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m) \leq \widehat{c}^B_{1-\alpha,N}, \widehat{c}^B_{1-\alpha,N} \leq \mathrm{SW}_p^p(\mu,\nu) - \gamma(n,p,\beta,D,L)\right) \\
&\leq \mathbb{P}_{\mu\times\nu\times r\times\sigma}\left(\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m) - \mathrm{SW}_p^p(\mu,\nu) \leq -\gamma(n,p,\beta,D,L)\right) \\
&\leq \mathbb{P}_{\mu\times\nu\times r\times\sigma}\left(\left|\widehat{\mathrm{SW}}_p^p(\widehat{\mu}_n,\widehat{\nu}_m) - \mathrm{SW}_p^p(\mu,\nu)\right| \geq \gamma(n,p,\beta,D,L)\right) \\
&\leq \frac{\beta}{2}.
\end{aligned}$$

To conclude, note that whenever $\mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{B}) > 1 - \frac{\beta}{2}$, the theorem of total probability implies

$$\begin{aligned}
\mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{A}) &= \mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{A}\cap\mathcal{B}) + \mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{A}\cap\mathcal{B}^c) \\
&= \mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{A}\cap\mathcal{B}) + \mathbb{P}_{p\times q\times r\times\sigma}(\mathcal{A} \mid \mathcal{B}^c)\,\mathbb{P}_{\mu\times\nu\times r\times\sigma}(\mathcal{B}^c) \\
&\leq \frac{\beta}{2} + \frac{\beta}{2}\cdot 1 \\
&= \beta.
\end{aligned}$$

■

The condition for achieving the desired test power, given in Lemma 19, involves two main components.: the term $\gamma(n,p,\beta,D,L)$, which corresponds to the sample complexity bound (see Lemma 18), and the random empirical quantile threshold $\widehat{c}^B_{1-\alpha,N}$, which constitutes the main difficulty in the theoretical analysis of permutation-based testing. To address this issue, we first rely on Domingo-Enrich et al. (2023, Lemma 6), which establishes a connection between the (random) empirical quantile threshold $\widehat{c}^B_{1-\alpha,N}$ and the (deterministic) quantile $c_{1-\alpha,N}$ of the permuted test statistic, conditional on the samples and the projection directions.

**Lemma 20 (Relation between population and empirical quantiles)** *Let $\alpha,\beta \in (0,1)$, and let $N \geq 1$. Denote by $B$ the number of sampled permutations as described in Algorithm 1. Given the datasets $\mathcal{Y}_n = (y_1,\ldots,y_n)$ and $\mathcal{Z}_m = (z_1,\ldots,z_m)$, whose points lie in the ball centered at the origin with radius $D$, as well as a set of projection directions $\Theta = (\theta_1,\ldots,\theta_L)$ on the unit sphere $\mathbb{S}^{d-1}$, we have*

$$\mathbb{P}_r\big(\widehat{c}^B_{1-\alpha,N} \leq c_{1-\alpha_1,N} \;\big|\; \mathcal{Y}_n, \mathcal{Z}_m, \Theta\big) > 1 - \tfrac{\beta}{2},$$

*where* $\alpha_1 \coloneqq \left( \dfrac{\beta/2}{\binom{B}{\lfloor \alpha(B+1)\rfloor}} \right)^{1/\lfloor \alpha(B+1)\rfloor}$.

Consequently, obtaining an upper bound for $c_{1-\alpha_1,N}$ immediately implies an upper bound for the empirical quantile threshold $\widehat{c}^B_{1-\alpha,N}$. The next proposition establishes such an upper bound for the population quantile.

**Proposition 21 (Upper bound for $c_{1-\alpha,N}$)** *Let $p \geq 1$ and $\alpha \in (0,1)$. Under the same assumptions as in Lemma 20, and for brevity, we denote $N := n+m$. Moreover, without loss of generality, we assume that $n \leq m$. Then the $(1-\alpha)$-quantile of the permutation distribution, $c_{1-\alpha,N}$ introduced in (5), satisfies*

$$c_{1-\alpha,N} \;\leq\; \sqrt{\frac{8(2D)^{2p}\log(1/\alpha)}{3n}} + \frac{(4D)^p\sqrt{2}}{\sqrt{n}}. \tag{9}$$

**Proof** For readability, we denote

$$h(n,m) := \frac{2\left(n+m-\frac{1}{2}\right)}{nm\left(\frac{(2D)^p}{n}+\frac{(2D)^p}{m}\right)^2}\cdot\left(1-\frac{1}{2\max\{m,n\}}\right).$$

Setting $\epsilon := \sqrt{\ln\left(\frac{1}{\alpha}\right)\frac{1}{h(n,m)}}$, Lemma 12 guarantees

$$\mathbb{P}_\pi\left[\widehat{\mathrm{SW}}_p^{p,\pi} \geq \sqrt{\ln\left(\frac{1}{\alpha}\right)\frac{1}{h(n,m)}} + \mathbb{E}_\pi(\widehat{\mathrm{SW}}_p^{p,\pi})\,\Big|\,\mathcal{Y}_n,\mathcal{Z}_m,\Theta\right] \leq \alpha.$$

It implies, by definition of $c_{1-\alpha,N}$ (recalled in Eq. (6)), that

$$c_{1-\alpha,N} \leq \sqrt{\ln\left(\frac{1}{\alpha}\right)\cdot\frac{1}{h(n,m)}} + \mathbb{E}_\pi(\widehat{\mathrm{SW}}_p^{p,\pi}). \tag{10}$$

We now derive a more explicit and tractable expression for $h(n,m)$. Since $1 \leq n \leq m$, we have

$$h(n,m) = \frac{n(2m-1)(n+m-\frac{1}{2})}{(2D)^{2p}(n+m)^2}.$$

Setting $x := \dfrac{m}{n} \geq 1$, we obtain

$$\begin{aligned} h(n,m) &= \frac{n\left(2x-\frac{1}{n}\right)\left(1+x-\frac{1}{2n}\right)}{(2D)^{2p}(1+x)^2} \\ &\geq \frac{n\left(2x-\frac{1}{n}\right)\left(1+x-\frac{1}{2n}\right)}{(2D)^{2p}(1+x)^2}, \end{aligned}$$

where the last inequality follows from the facts that $2x-\dfrac{1}{n} \geq 2x-1$ and $1+x-\dfrac{1}{2n} \geq x+\dfrac{1}{2}$.

Moreover, since the function $\phi(x) = \dfrac{(2x-1)\left(x+\frac{1}{2}\right)}{(1+x)^2}$ is increasing for $x \geq 1$, it follows that

$$h(n,m) \geq \frac{3n}{8(2D)^{2p}}. \tag{11}$$

In Proposition 17, we have shown that, given the samples $\mathcal{Y}_n$, $\mathcal{Z}_m$ and the projection directions $\Theta = (\theta_1,\ldots\theta_L)$, the following bound holds:

$$\mathbb{E}_\pi\left[\widehat{\mathrm{SW}}_p^{p,\pi}\right] \leq (4D)^p\frac{\sqrt{2}}{\sqrt{n}}. \tag{12}$$

Combining inequalities (10),(11), and (12), we obtain

$$c_{1-\alpha,N} \;\leq\; \sqrt{\frac{8(2D)^{2p}\log(1/\alpha)}{3n}} + \frac{(4D)^p\sqrt{2}}{\sqrt{n}}.$$

■

Having established an upper bound for $c_{1-\alpha,N}$, we now rely on Lemma 20 to derive the following proposition, which provides a bound for the empirical quantile $\widehat{c}^B_{1-\alpha,N}$.

**Proposition 22 (Bound on the empirical quantile)** *Let $\beta > 0$ and $\frac{1}{B+1} \leq \alpha < 1$, where $B$ denotes the number of permutations as described in Algorithm 1. Set $\omega_\alpha \coloneqq \lfloor \alpha(B+1) \rfloor$. Then, conditional on the samples $\mathcal{Y}_n$ and $\mathcal{Z}_m$ (with $n \leq m$ and $N \coloneqq n+m$) whose points lie in the ball centered at the origin with radius $D$, as well as on a set of projection directions $\Theta = (\theta_1, \ldots, \theta_L)$ on the unit sphere $\mathbb{S}^{d-1}$, we have*

$$\mathbb{P}_r\left(\widehat{c}^B_{1-\alpha,N} \;\leq\; \sqrt{\frac{8(2D)^{2p}}{3n}\,\log\left(\frac{2e}{\alpha\left(\frac{\beta}{2}\right)^{1/\omega_\alpha}}\right)} \;+\; \frac{(4D)^p}{\sqrt{n}}\right) \;>\; 1-\frac{\beta}{2}.$$

**Proof** Proposition 21 guarantees that

$$c_{1-\alpha_1,N} \;\leq\; \sqrt{\frac{8(2D)^{2p}\log(1/\alpha_1)}{3n}} + \frac{(4D)^p\sqrt{2}}{\sqrt{n}}, \tag{13}$$

where

$$\alpha_1 = \left(\frac{\beta/2}{\binom{B}{\lfloor \alpha(B+1)\rfloor}}\right)^{1/\lfloor \alpha(B+1)\rfloor}.$$

To make this bound more explicit, we upper bound the logarithmic term. Let $\omega_\alpha \coloneqq \lfloor \alpha(B+1)\rfloor \geq 1$. Then

$$\log\left(\binom{B}{\omega_\alpha}^{1/\omega_\alpha}\right) \leq \log\left(\frac{eB}{\omega_\alpha}\right) \leq \log\left(\frac{2eB}{\alpha(B+1)}\right) \leq \log\left(\frac{2e}{\alpha}\right). \tag{14}$$

The first inequality follows from the bound $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$, and the second from the fact that $\lfloor x \rfloor \geq \frac{x}{2}$ whenever $x \geq 1$.

Combining Eq. (13) and Eq. (14) yields

$$c_{1-\alpha_1,N} \;\leq\; \sqrt{\frac{8(2D)^{2p}}{3n}\,\log\left(\frac{2e}{\alpha\left(\frac{\beta}{2}\right)^{1/\omega_\alpha}}\right)} \;+\; \frac{(4D)^p\sqrt{2}}{\sqrt{n}}. \tag{15}$$

Hence, from Lemma 20, Eq. (15), and also conditioning on the samples $\mathcal{Y}_n$, $\mathcal{Z}_m$, and the projection directions $\theta_1, \ldots, \theta_L$, we have

$$\mathbb{P}_r\left(\widehat{c}^B_{1-\alpha,N} \;\leq\; \sqrt{\frac{8(2D)^{2p}}{3n}\,\log\left(\frac{2e}{\alpha\left(\frac{\beta}{2}\right)^{1/\omega_\alpha}}\right)} \;+\; \frac{(4D)^p\sqrt{2}}{\sqrt{n}}\right) \;>\; 1-\frac{\beta}{2}.$$

■

Finally, Theorem 5 is proved by combining Lemma 19 and Proposition 22, as outlined in the following.

**Proof of Theorem 5**

To ensure that the test has power at least $1-\beta$, Lemma 19 shows it suffices to find conditions on $\mu$ and $\nu$ such that

$$\mathbb{P}_{\mu\times\nu\times r\times\sigma}\left[\mathrm{SW}_p^p(\mu,\nu) \;\geq\; (2D)^p\left(\sqrt{\frac{\log(8/\beta)}{2L}}+\sqrt{\frac{\log(8/\beta)}{n}}\right)+\frac{C_{p,D}}{\sqrt{n}}+\widehat{c}^{B}_{1-\alpha,N}\right] > 1-\frac{\beta}{2}, \tag{16}$$

where the constant $C_{p,D}$ is introduced in Lemma 18.

From Proposition 22 and the condition provided in Eq. (16), a sufficient condition for guaranteeing a power at least $1-\beta$ is given by

$$\sqrt{\frac{8(2D)^{2p}}{3n}\log\left(\frac{2e}{\alpha\left(\frac{\beta}{2}\right)^{1/\omega_\alpha}}\right)} + \frac{(4D)^p\sqrt{2}}{\sqrt{n}} \leq -(2D)^p\left(\sqrt{\frac{\log(8/\beta)}{2L}}+\sqrt{\frac{\log(8/\beta)}{n}}\right)-\frac{C_{p,D}}{\sqrt{n}}+\mathrm{SW}_p^p(\mu,\nu)$$

or equivalently,

$$\mathrm{SW}_p^p(\mu,\nu) \geq \sqrt{\frac{8(2D)^{2p}}{3n}\log\left(\frac{2e}{\alpha\left(\frac{\beta}{2}\right)^{1/\omega_\alpha}}\right)} + \frac{(4D)^p\sqrt{2}}{\sqrt{n}}+(2D)^p\left(\sqrt{\frac{\log(8/\beta)}{2L}}+\sqrt{\frac{\log(8/\beta)}{n}}\right)+\frac{C_{p,D}}{\sqrt{n}}.$$

The above inequality implies the claimed bound on the test's power.

# 12 LOWER BOUNDS

In this section, we present the proofs of the lower bounds stated in the main body, namely Proposition 6 and Proposition 7. We begin by explaining, in Section 12.1, why the problem of obtaining lower bounds for two-sample testing can be reduced to that of one-sample testing—i.e., goodness-of-fit testing. Next, we present the proofs of the two propositions.

## 12.1 Reduction to goodness-of-fit testing

Following Arias-Castro et al. (2018, Lemma 1), we recast the task of establishing lower bounds for the two-sample testing problem as the conceptually simpler task of establishing lower bounds for the one-sample testing problem (also called goodness-of-fit). Intuitively, goodness-of-fit can be viewed as a special case of two-sample testing in which one of the distributions is fully known—which is equivalent to having access to an infinite number of samples from it.

Fix a known reference distribution $\nu_0$ and observe independent samples $Y_1,\ldots,Y_n\sim\nu$, where $\nu$ is an unknown distribution. We consider the following goodness-of-fit testing problem:

$$\mathcal{H}_0:\ \nu=\nu_0 \qquad \text{vs.} \qquad \mathcal{H}_1:\ \mathrm{SW}_p^p(\nu,\nu_0)\ \geq \epsilon_n.$$

Similarly to two-sample testing (see Section 2.3), we define the minimax separation for one-sample testing at level $\alpha$ and power $1-\beta$ as

$$\epsilon_n^\dagger := \inf\left\{\epsilon_n>0: \inf_{\Delta_n\in\Phi_{\alpha,n,\nu_0}}\ \sup_{\nu\in\mathcal{P}_1(\epsilon_n,\nu_0)} \mathbb{E}_\nu[1-\Delta_n]\leq\beta\right\},$$

where

- $\Phi_{\alpha,n,\nu_0}$ is the set of tests $\Delta_n$, based on $n$ samples, of level at most $\alpha$, that is, $\mathbb{E}_{\nu_0}[\Delta_n]\leq\alpha$;
- $\mathcal{P}_1(\epsilon_n,\nu_0)$ is the set of distributions $\nu$ such that $\mathrm{SW}_p^p(\nu,\nu_0)\ \geq\ \epsilon_n$.

Because the two-sample setting involves estimating both distributions from data—unlike the one-sample case, where one distribution is known exactly—the minimax separation for two-sample testing cannot be smaller than that for one-sample testing. Consequently,

$$\epsilon^{\dagger}_{n,m} \;\geq\; \epsilon^{\dagger}_{n},$$

where $\epsilon^{\dagger}_{n}$ and $\epsilon^{\dagger}_{n,m}$ denote the minimax separations for the one-sample and two-sample problems, respectively. In particular, a lower bound on the minimax separation in the one-sample setting directly implies a lower bound in the two-sample setting.

In what follows, we prove that $\epsilon^{\dagger}_{n} \gtrsim n^{-1/2}$ which immediately yields $\epsilon^{\dagger}_{n,m} \gtrsim n^{-1/2}$.

### 12.2 Proof of Proposition 6

Let $[d] \coloneqq \{1, \ldots, d\}$, and denote by $\mathcal{P}^{(d)}_{\text{Multi}}$ the class of multinomial distributions on $[d]$. In this section, we provide the proof of Proposition 6, which establishes a lower bound on the minimax separation rate for the two-sample testing problem over $\mathcal{P}^{(d)}_{\text{Multi}} \times \mathcal{P}^{(d)}_{\text{Multi}}$.

The proof of Proposition 6 is based on Ingster's method (Ingster, 1993), a classical approach for deriving minimax lower bounds in hypothesis testing. We first provide a brief overview of this technique.

**Ingster's method for minimax lower bounds.** Let $\nu_0$ denote a fixed null distribution. The minimax Type II error at level $\alpha$ (for one-sample testing problem) is defined as

$$R^{\dagger}_{n,\alpha} \coloneqq \inf_{\phi \in \Phi_{\alpha,n,\nu_0}} \; \sup_{\nu \in \mathcal{P}_1(\epsilon_n, \nu_0)} \mathbb{P}_{\nu}\big(\phi = 0\big),$$

where $\Phi_{\alpha,n,\nu_0}$ and $\mathcal{P}_1(\epsilon_n, \nu_0)$ are defined in Section 12.1.

Select $T \geq 1$ distributions $\nu_1, \ldots, \nu_T \in \mathcal{P}_1(\epsilon_n, \nu_0)$, and define a mixture distribution $Q$ such that, for any measurable set $A$,

$$Q(A) \coloneqq \frac{1}{T} \sum_{t=1}^{T} \nu_t^{\otimes n}(A),$$

where the $n$-fold product distribution of a distribution $\nu_t$ is denoted as $\nu_t^{\otimes n}$.

Given $n$ i.i.d. observations $Y_1, \ldots, Y_n$ drawn from an unknown distribution $\nu$, we denote the likelihood ratio between $Q$ and the null distribution $\nu_0$ by

$$L_n(Y_1, \ldots, Y_n) \coloneqq \frac{\mathrm{d}Q}{\mathrm{d}\nu_0^{\otimes n}}(Y_1, \ldots, Y_n) \;=\; \frac{1}{T} \sum_{t=1}^{T} \prod_{i=1}^{n} \frac{\nu_t(Y_i)}{\nu_0(Y_i)},$$

where, with a slight abuse of notation, we denote by $\nu_0$ and $(\nu_t)_{t=1}^{T}$ both the probability measures and their densities with respect to the Lebesgue measure.

Given these definitions, the following lemma provides a condition on the likelihood ratio between the mixture distribution $Q$ and the null distribution $\nu_0$ to obtain a lower bound for the minimax Type II error $R^{\dagger}_{n,\alpha}$.

**Lemma 23 (Lower bound via a mixture)** *Fix $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$. If*

$$\mathbb{E}_{\nu_0}\big[L_n^2\big] \;\leq\; 1 + 4(1-\alpha-\beta)^2,$$

*then $R^{\dagger}_{n,\alpha} \;\geq\; \beta$.*

The proof of this lemma can be found in Wasserman (2014, Section 11) and in Kim et al. (2022, Appendix H).

To apply Lemma 23, we need to (i) specify the distributions $Q$ and $\nu_0$, (ii) compute the expectation of the squared likelihood ratio under the null. We borrow the construction of perturbed distributions around the

uniform measure on $[d]$ from Kim et al. (2022, Appendix H), and adapt the corresponding computations to the sliced Wasserstein distance setting.

Let $\nu_0$ denote the uniform distribution over the set $[d] \coloneqq \{1, \ldots, d\}$, that is, $\nu_0(k) = \frac{1}{d}$ for $k = 1, \ldots, d$.

Define the set

$$\mathcal{M}_d \coloneqq \{\eta \in \{-1, 1\}^d : \sum_{k=1}^{d} \eta_k = 0\}. \tag{17}$$

We may assume without loss of generality that $d$ is even, as the proof for the odd-dimensional case follows analogously by setting the last coordinate to zero $\eta_d = 0$. Given $\eta \in \mathcal{M}_d$, we define the measure $\nu_\eta$ as

$$\nu_\eta(k) \coloneqq \nu(k) + \frac{2\epsilon_n}{d}\eta_k, \quad \text{for } k = 1, \ldots, d,$$

for some $\epsilon_n \leq \frac{1}{2}$ that we will specify later in the proof.

Note that, for any $\eta \in \mathcal{M}_d$, the distribution $\nu_\eta$ is a valid probability distribution on the set $[d]$. Indeed,

1. $\nu_\eta$ is non negative.
2. $\displaystyle\sum_{k=1}^{d} \nu_\eta(k) = \sum_{k=1}^{d} \left(\nu(k) + \frac{2\epsilon_n}{d}\eta_k\right) = \sum_{k=1}^{d} \nu(k) + \frac{2\epsilon_n}{d}\sum_{k=1}^{d} \eta_k = 1.$

Let us now evaluate the sliced Wasserstein distance between $\nu_0$ and its perturbation $\nu_\eta$. By construction, exactly $d/2$ coordinates $k$ satisfy $\eta_k = 1$, and the remaining $d/2$ coordinates satisfy $\eta_k = -1$. For coordinates with $\eta_k = 1$, the weight in $\nu_\eta$ is increased by $\frac{2\epsilon_n}{d}$ compared to $\nu_0$, while for coordinates with $\eta_k = -1$, the weight is decreased by the same amount. Hence, the total excess mass between the $d/2$ heavier points and the $d/2$ lighter points is

$$\frac{d}{2} \cdot \frac{2\epsilon_n}{d} = \epsilon_n.$$

Since $\nu_0$ is uniform over $[d]$ and the distance between any two distinct points in $[d]$ is at least 1, moving any unit of excess mass incurs a cost of at least $1^2 = 1$. Therefore, the total transportation cost is at least $\epsilon_n$. Since both $\nu_0$ and $\nu_\eta$ are one-dimensional distributions, the sliced Wasserstein distance coincides with the standard Wasserstein distance and

$$\mathrm{SW}_p^p(\nu_0, \nu_\eta) = \mathrm{W}_p^p(\nu_0, \nu_\eta) \geq \epsilon_n.$$

Let $T$ be the cardinality of $\mathcal{M}_d$. In order to apply Ingster's method, we introduce the uniform mixture $\nu_{\eta(1)}, \ldots, \nu_{\eta(T)}$,

$$Q \coloneqq \frac{1}{T} \sum_{\eta \in \mathcal{M}_d} \nu_\eta.$$

The likelihood ratio between $Q$ and the null distribution $\nu_0$ will be

$$L_n(Y_1, \ldots, Y_n) \coloneqq \frac{1}{T} \sum_{\eta \in \mathcal{M}_d} \prod_{i=1}^{n} \frac{\nu_\eta(Y_i)}{\nu_0(Y_i)}.$$

With those ingredients on hand, we are now ready to compute the expected value of the squared likelihood ratio. The subsequent computation follows exactly the same steps as in Kim et al. (2022, Appendix H). We present it here for completeness.

$$L_n^2(Y_1, \ldots, Y_n) = \frac{1}{T^2} \sum_{\eta, \eta' \in \mathcal{M}_d} \prod_{i=1}^{n} \frac{\nu_\eta(Y_i)\, \nu_{\eta'}(Y_i)}{\nu_0(Y_i)^2}$$

$$= \frac{1}{T^2} \sum_{\eta,\eta' \in \mathcal{M}_d} \prod_{i=1}^{n} \frac{(\frac{1}{d} + \frac{2\epsilon_n}{d}\eta_{Y_i})(\frac{1}{d} + \frac{2\epsilon_n}{d}\eta'_{Y_i})}{1/d^2}$$

$$= \frac{1}{T^2} \sum_{\eta,\eta' \in \mathcal{M}_d} \prod_{i=1}^{n} (1 + 2\epsilon_n \eta_{Y_i})(1 + 2\epsilon_n \eta'_{Y_i}).$$

By definition of $\eta \in \mathcal{M}_d$, taking the expectation over $Y_1, \ldots, Y_n \sim \nu_0$, we obtain

$$\mathbb{E}_{\nu_0}\left[L_n^2\right] = \frac{1}{T^2} \sum_{\eta,\eta' \in \mathcal{M}_d} \left(1 + \frac{4\epsilon_n^2}{d} \sum_{k=1}^{d} \eta_k \eta'_k\right)^n.$$

Moreover, using the inequality $1 + x \leq e^x$, which holds for any $x \in \mathbb{R}$,

$$\mathbb{E}_{\nu_0}\left[L_n^2\right] \leq \frac{1}{T^2} \sum_{\eta,\eta' \in \mathcal{M}_d} \exp\left(\frac{4n\epsilon_n^2}{d} \sum_{k=1}^{d} \eta_k \eta'_k\right).$$

Let $\eta$ and $\eta^*$ be independent random variables uniformly distributed on $\mathcal{M}_d$. We have

$$\frac{1}{T^2} \sum_{\eta,\eta' \in \mathcal{M}_d} \exp\left(\frac{4n\epsilon_n^2}{d} \sum_{k=1}^{d} \eta_k \eta'_k\right) = \mathbb{E}_{\eta,\eta^*}\left[\exp\left(\frac{4n\epsilon_n^2}{d}\langle \eta, \eta^* \rangle\right)\right].$$

Applying Lemma 2 in Dubhashi and Ranjan (1996), we obtain

$$\begin{aligned}
\mathbb{E}_{\nu_0}\left[L_n^2\right] &\leq \mathbb{E}_{\eta,\eta^*}\left[\exp\left(\frac{4n\epsilon_n^2}{d}\langle \eta, \eta^* \rangle\right)\right] \\
&\leq \prod_{k=1}^{d} \mathbb{E}_{\eta_k,\eta_k^*}\left[\exp\left(\frac{4n\epsilon_n^2}{d}\eta_k \eta_k^*\right)\right] \\
&= \prod_{k=1}^{d} \cosh\left(\frac{4n\epsilon_n^2}{d}\right) \\
&\leq \exp\left(\frac{d}{2}\left(\frac{4n\epsilon_n^2}{d}\right)^2\right) = \exp\left(\frac{8n^2\epsilon_n^4}{d}\right),
\end{aligned}$$

where the last inequality follows from $\cosh(x) \leq e^{x^2/2}$, which holds for any $x \in \mathbb{R}$.

The proof is concluded by observing that Lemma 23 guarantees that the minimax Type II error is lower bounded by $\beta$ provided that

$$\epsilon_n \leq \frac{1}{\sqrt{n}} \sqrt[4]{\frac{d \log\left[1 + 4(1 - \alpha - \beta)^2\right]}{8}}.$$

### 12.3 Proof of Proposition 7

Proposition 6 states a lower bound on the minimax separation rate for the two-sample testing problem over the class of multinomial distributions. Turning to the more general class of distributions with bounded support, denoted by $\mathcal{P}_{\mathbb{R}^d}(D)$ for some $D > 0$, the proof of Proposition 7 establishes the corresponding lower bound on the minimax separation rate for the two-sample testing problem over this class.

We begin by presenting some preliminary ingredients required for the proof of Proposition 7.

#### Main ingredients for the proof

We begin this section by recalling a basic fact: the uniform distribution on the sphere is invariant under orthogonal transformations, as stated below.

**Lemma 24 (Rotation invariance)** *If $U$ is uniform on $\mathbb{S}^{d-1}$ and $Q$ is an orthogonal matrix, then $QU$ has the same distribution as $U$.*

**Proof** Let $\sigma$ denote the $(d-1)$-dimensional surface area measure on $\mathbb{S}^{d-1}$. Orthogonal maps preserve surface area on the sphere, that is, $\sigma(QA) = \sigma(A)$ for every Borel set $A \subset \mathbb{S}^{d-1}$. Hence, for any Borel set $A$, we have

$$\mathbb{P}(QU \in A) = \mathbb{P}(U \in Q^{-1}A) = \frac{\sigma(Q^{-1}A)}{\sigma(\mathbb{S}^{d-1})} = \frac{\sigma A}{\sigma(\mathbb{S}^{d-1})} = \mathbb{P}(U \in A).$$

Therefore $QU$ and $U$ have the same law. ■

We now identify the law of a single coordinate under the uniform distribution on the sphere.

**Lemma 25 (First coordinate of a spherical uniform distribution)** *Let $U = (U_1, \dots, U_d) \sim \mathrm{Unif}(\mathbb{S}^{d-1})$.*

1. *For $d = 1$, $U_1^2 \equiv 1$ a.s. (a degenerate distribution at 1).*
2. *For $d \geq 2$, $U_1^2 \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$.*

**Proof**

**Case $d = 1$.** It is clear that $\mathbb{S}^0 = \{\pm 1\}$ and $U_1 = \pm 1$ with probability 1/2 each. Therefore, $U_1^2 \equiv 1$ almost surely.

**Case $d \geq 2$.** Let $Z = (Z_1, \dots, Z_d)$ be standard normal in $\mathbb{R}^d$ (i.e., $Z \sim \mathcal{N}(0, I_d)$). By Equation (3.15) in Vershynin (2018), $U := Z/\|Z\| \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Since $Z_1 \sim \mathcal{N}(0,1)$ and the $Z_i$'s are independent,

$$X := Z_1^2 \sim \chi_1^2 = \Gamma\left(\frac{1}{2}; 2\right), \qquad Y := \sum_{i=2}^{d} Z_i^2 \sim \chi_{d-1}^2 = \Gamma\left(\frac{d-1}{2}; 2\right),$$

with $X \perp Y$.

Finally, using the classical fact that if $M \sim \chi^2(m)$ and $N \sim \chi^2(n)$ are independent, then

$$\frac{M}{M+N} \sim \mathrm{Beta}\left(\frac{m}{2}, \frac{n}{2}\right),$$

(see, e.g., Soch et al. (2025, Section 3.9)), we conclude that

$$U_1^2 \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right).$$

As a consequence, for $p \geq 1$, and following the result of Balakrishnan and Nevzorov (2004, Page 145), we have

$$\mathbb{E}\,|U_1|^p = \begin{cases} \dfrac{\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p+d}{2}\right)}, & d \geq 2, \\ 1, & d = 1, \end{cases}$$

where $\Gamma(\cdot)$ denotes the classical Gamma function. ■

We recall two classical inequalities that are particularly useful for bounding Gamma functions: Wendel's inequality and Stirling's inequalities.

**Lemma 26 (Wendel's inequality)** *We denote by $\Gamma(\cdot)$ the classical Gamma function. For $0 \leq a < 1$ and $x > 0$,*

$$\left(\frac{x}{x+a}\right)^{1-a} \leq \frac{\Gamma(x+a)}{x^a \Gamma(x)} \leq 1.$$

**Lemma 27 (Stirling's inequality)** *We denote by $\Gamma(\cdot)$ the classical Gamma function. For $x > 0$,*

$$\sqrt{2\pi}\, x^{x-\frac{1}{2}} e^{-x} \;\le\; \Gamma(x) \;\le\; \sqrt{2\pi}\, x^{x-\frac{1}{2}} e^{-x} e^{\frac{1}{12x}}.$$

Proofs of Lemmas 26 and 27 can be found in Wendel (1948) and Jameson (2015, Theorem 1), respectively.

We then close these preliminaries with a simple yet essential fact from linear algebra.

**Lemma 28** *For any unit vector $u \in \mathbb{R}^d$ with $\|u\| = 1$, there exists a rotation $Q \in SO(d)$ such that $Qe_1 = u$, where $e_1 = (1, 0, \ldots, 0)^\top$ and $SO(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = I,\ \det Q = 1\}$.*

**Proof** Extend $\{u\}$ to a basis of $\mathbb{R}^d$ and apply the Gram–Schmidt procedure to obtain an orthonormal basis $\{u, v_2, \ldots, v_d\}$. Let

$$Q \coloneqq [\, u \mid v_2 \mid \cdots \mid v_d \,] \in \mathbb{R}^{d\times d}.$$

Then the columns of $Q$ are orthonormal and $Q^\top Q = I$. In particular, $Q$ is orthogonal and satisfies $Qe_1 = u$.

If $\det(Q) = 1$, then $Q \in SO(d)$ and we are done. If $\det(Q) = -1$, replace one column $v_j$ with $-v_j$ for some $j \ge 2$ (e.g., $j = 2$). The resulting matrix

$$\widetilde{Q} = [\, u \mid -v_2 \mid v_3 \mid \cdots \mid v_d \,]$$

remains orthogonal, still satisfies $\widetilde{Q}e_1 = u$, and has $\det(\widetilde{Q}) = -\det(Q) = 1$. Therefore, there exists $Q \in SO(d)$ such that $Qe_1 = u$. ■

With these preparatory results in place, we are now ready to prove Proposition 7.

### 12.3.1 Formal proof

As discussed in Section 12.1, once a lower bound for the minimax rate $\epsilon_n^\dagger$ for goodness-of-fit testing is established, a lower bound $\epsilon_{n,m}^\dagger$ for two-sample testing follows immediately. For the general case of distributions with bounded support, we apply Le Cam two-point method (Le Cam, 1973, 2012), a classical techniques for deriving minimax lower bounds.

Let $\nu_0$ be a reference distribution. Let $\epsilon_n > 0$ and let $\mu_0$ be an alternative distribution in $\mathcal{P}_1(\epsilon_n, \nu_0)$. Using the same notation as for the one-sample testing problem (see Section 12.1), we obtain

$$\begin{aligned}
\inf_{\Delta_n \in \Phi_{\alpha,n,\nu_0}} \sup_{\mu \in \mathcal{P}_1(\epsilon_n,\nu_0)} \mathbb{E}_\mu[1-\Delta_n] &\ge \inf_{\Delta_n \in \Phi_{\alpha,n,\nu_0}} \mathbb{E}_{\mu_0}[1-\Delta_n] \\
&= 1 - \sup_{\Delta_n \in \Phi_{\alpha,n,\nu_0}} \mathbb{E}_{\mu_0}[\Delta_n] \\
&= 1 - \sup_{\Delta_n \in \Phi_{\alpha,n,\nu_0}} \{\mathbb{E}_{\mu_0}[\Delta_n] - \mathbb{E}_{\nu_0}[\Delta_n] + \mathbb{E}_{\nu_0}[\Delta_n]\} \\
&\ge 1 - \alpha - \sup_{\Delta_n \in \Phi_{\alpha,n,\nu_0}} \{\mathbb{E}_{\mu_0}[\Delta_n] - \mathbb{E}_{\nu_0}[\Delta_n]\} \\
&\ge 1 - \alpha - d_{\mathrm{TV}}\big(\mu_0^{\otimes n}, \nu_0^{\otimes n}\big) \\
&\overset{(i)}{\ge} 1 - \alpha - 1 + \tfrac{1}{2} e^{-D_{\mathrm{KL}}\left(\mu_0^{\otimes n} \,\|\, \nu_0^{\otimes n}\right)} \\
&\overset{(ii)}{=} \tfrac{1}{2} e^{-n\, D_{\mathrm{KL}}(\mu_0 \,\|\, \nu_0)} - \alpha,
\end{aligned}$$

where $(i)$ follows from the Bretagnolle–Huber inequality introduced in Canonne (2022, Lemma 3), and $(ii)$ uses the chain rule for the KL divergence. We recall that $\mu_0^{\otimes n}$ and $\nu_0^{\otimes n}$ are the $n$-fold product distributions of $\mu_0$ and $\nu_0$, respectively.

Consequently, the minimax Type II error is at least $\beta$, that is,

$$\inf_{\Delta_n \in \Phi_{\alpha,n,\nu_0}} \sup_{\mu \in \mathcal{P}_1(\epsilon_n,\nu_0)} \mathbb{E}_\mu[1-\Delta_n] \ge \beta,$$

provided that

$$\alpha + \beta < 0.5 \quad \text{and} \quad D_{\mathrm{KL}}(\mu_0 \,\|\, \nu_0) \;\le\; \frac{1}{n} \log\!\left(\frac{1}{2(\alpha+\beta)}\right).$$

In order to obtain a lower bound, we need to choose $\mu_0, \nu_0 \in \mathcal{P}_{\mathbb{R}^d}(D)$ such that

$$D_{\mathrm{KL}}(\mu_0 \,\|\, \nu_0) \;\leq\; \frac{1}{n}\log\left(\frac{1}{2(\alpha+\beta)}\right). \tag{18}$$

and

$$\mathrm{SW}_p^p(\mu_0, \nu_0) \geq \epsilon_n. \tag{19}$$

Inspired by the works of Kim and Schrab (2023, Section E.10.1) and adapting their work to the case of the sliced Wasserstein distance, we choose

$$\mu_0 = p_0\delta_x + (1-p_0)\delta_v, \quad \nu_0 = q_0\delta_x + (1-q_0)\delta_v,$$

where $x, v \in \mathbb{R}^d$ with $\|x-v\|_2 = D$, $0 \leq p_0, q_0 \leq 1$, and $\delta_x$ denotes the Dirac measure at $x$. Furthermore, we also choose

$$p_0 = \frac{1}{2} + \min\left\{\sqrt{\frac{1}{4n}\log\left(\frac{1}{2(\alpha+\beta)}\right)},\, \frac{1}{2}\right\}, \quad q_0 = \frac{1}{2}.$$

With those parameters, we have

$$D_{\mathrm{KL}}(\mu_0 \,\|\, \nu_0) \leq \frac{(p_0-q_0)^2}{q_0(1-q_0)} = \min\left\{\frac{1}{n}\log\left(\frac{1}{2(\alpha+\beta)}\right), 1\right\} \leq \frac{1}{n}\log\left(\frac{1}{2(\alpha+\beta)}\right),$$

where the first inequality follows directly from a classical inequality between the KL and $\chi^2$ divergences (see, e.g., Tsybakov (2008, Lemma 2.7)).

Moreover, we have

$$\begin{aligned}
\mathrm{SW}_p^p(\mu_0, \nu_0) &= \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p\Big(\Pi_\#^\theta \mu_0,\, \Pi_\#^\theta \nu_0\Big)\, \sigma(d\theta) \\
&\overset{(i)}{=} \int_{\mathbb{S}^{d-1}} |p_0 - q_0|\, |\langle x-v, \theta\rangle|^p\, \sigma(d\theta) \\
&\overset{(ii)}{=} |p_0 - q_0|\, \|x-v\|^p \int_{\mathbb{S}^{d-1}} |\langle u, \theta\rangle|^p\, \sigma(d\theta),
\end{aligned}$$

where (i) follows from the computation of the Wasserstein distance between two Dirac measures in one dimension (see, e.g., Chewi et al. (2025, Example 1.4)), and (ii) from setting $u = \frac{x-v}{\|x-v\|}$.

We can further develop this expression as

$$\begin{aligned}
\mathrm{SW}_p^p(\mu_0, \nu_0) &= |p_0 - q_0|\, \|x-v\|^p \mathbb{E}_{\theta\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[|\langle u, \theta\rangle|^p\right] \\
&\overset{(iii)}{=} |p_0 - q_0|\, \|x-v\|^p \mathbb{E}_{\theta\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[|\langle Qe_1, \theta\rangle|^p\right] \\
&= |p_0 - q_0|\, \|x-v\|^p \mathbb{E}_{\theta\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left|\langle e_1, Q^\top\theta\rangle\right|^p\right] \\
&\overset{(iv)}{=} |p_0 - q_0|\, \|x-v\|^p \mathbb{E}_{\theta\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[|\langle e_1, \theta\rangle|^p\right] \\
&= |p_0 - q_0|\, \|x-v\|^p \mathbb{E}_{\theta\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[|\theta_1|^p\right] \\
&\overset{(v)}{=} |p_0 - q_0|\, \|x-v\|^p \begin{cases} \dfrac{\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p+d}{2}\right)}, & d \geq 2, \\ 1, & d = 1, \end{cases}
\end{aligned}$$

where $(iii)$ follows from Lemma 28, $(iv)$ from Lemma 24, and $(v)$ from Lemma 25.

We now derive lower bound for the term

$$c_{d,p} \coloneqq \frac{\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p+d}{2}\right)}.$$

The term $\Gamma\left(\frac{p+1}{2}\right)$ can be controled by applying Stirling's lower bound (recalled in Lemma 27). In particular, we have

$$\Gamma\left(\frac{p+1}{2}\right) \;\geq\; \sqrt{2\pi}\left(\frac{p+1}{2}\right)^{\frac{p}{2}} e^{-\frac{(p+1)}{2}}.$$

We next control the remaining term.

Write $\frac{p}{2} = m + r$ with $m = \lfloor p/2 \rfloor \in \mathbb{N}$ and $r \in [0, 1)$. We then can decompose the remaining terms as follow:

$$\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+m+r)} = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+m)} \cdot \frac{\Gamma(\frac{d}{2}+m)}{\Gamma(\frac{d}{2}+m+r)}.$$

By the recursion formula of the gamma function, $\Gamma(p+1) = p\Gamma(p)$ for $p \geq 0$, we obtain

$$\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+m)} \;\geq\; (\tfrac{d}{2}+m)^{-m}.$$

For the second factor, since $0 \leq r < 1$, Wendel's inequality (recalled in Lemma 26) yields

$$\frac{\Gamma(\frac{d}{2}+m)}{\Gamma(\frac{d}{2}+m+r)} \;\geq\; \left(\tfrac{d}{2}+m+r\right)^{-r}.$$

Hence,

$$\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+p}{2})} = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+m+r)} \;\geq\; (\tfrac{d}{2}+m)^{-m}\,(\tfrac{d}{2}+m+r)^{-r} \;\geq\; \left(\tfrac{d+p}{2}\right)^{-\frac{p}{2}}.$$

Combining the above estimates gives

$$c_{d,p} \;\geq\; \frac{\sqrt{2\pi}}{\sqrt{\pi}}\left(\frac{p+1}{2}\right)^{\frac{p}{2}} e^{-\frac{(p+1)}{2}} \cdot \left(\frac{d+p}{2}\right)^{-\frac{p}{2}} = \sqrt{\frac{2}{e}}\left[\frac{p+1}{e(d+p)}\right]^{\frac{p}{2}}.$$

As a result, we have

$$\mathrm{SW}_p^p(\mu_0, \nu_0) \geq \frac{D^p}{\sqrt{2e}}\left[\frac{p+1}{e(d+p)}\right]^{\frac{p}{2}} \min\left\{\sqrt{\frac{1}{n}\log\left(\frac{1}{2(\alpha+\beta)}\right)},\; 1\right\}.$$

Setting $\epsilon_n := \frac{D^p}{\sqrt{2e}}\left[\frac{p+1}{e(d+p)}\right]^{\frac{p}{2}} \min\left\{\sqrt{\frac{1}{n}\log\left(\frac{1}{2(\alpha+\beta)}\right)},\; 1\right\}$, conditions (18) and (19) hold, and thus the minimax separation satisfies

$$\epsilon_n^{\dagger} \;\geq\; \epsilon_n = \frac{D^p}{\sqrt{2e}}\left[\frac{p+1}{e(d+p)}\right]^{\frac{p}{2}} \min\left\{\sqrt{\frac{1}{n}\log\left(\frac{1}{2(\alpha+\beta)}\right)},\; 1\right\}.$$

The above bound implies the bound stated in Proposition 7.

## 13 PERMUTATION APPROACH

In this section, we present the motivation for employing a permutation-based strategy in the two-sample testing problem using the sliced Wasserstein distance.

In hypothesis testing, the determination of the critical value plays a central role, as it directly governs the decision to reject the null hypothesis. In general, existing approaches for determining the critical value can be grouped into two main categories: asymptotic and non-asymptotic methods.

**Asymptotic method.** In the asymptotic approach, the critical value is determined from the limiting distribution of the test statistic under the null hypothesis (see, e.g., Shekhar et al. (2022, 2023); Zaremba et al. (2013)). Specifically, to control the Type I error in the asymptotic regime, the critical value is chosen as the $(1-\alpha)$ quantile of the null distribution.

However, this asymptotic approach is not without limitations. In many cases, the limiting null distribution is either intractable or lacks a convenient closed form. For instance, Gretton et al. (2012, Theorem 12) show that the null distribution of the squared empirical biased estimator of MMD is an infinite weighted sum of independent $\chi^2$ random variables, with weights given by the eigenvalues of the kernel operator. This makes the distribution highly dependent on the kernel choice and analytically difficult to handle. Likewise, more general tests based on U-statistics face similar challenges in the multinomial setting: the shape of the null distribution is intricately tied to the probabilities defining the underlying multinomial structure (see Kim et al. (2022, Figure 1)).

The test statistic $\widehat{\mathrm{SW}}_p^p$ defined in (2) faces the same challenge. To gain intuition about its null distribution, we approximate it empirically through simulation. Specifically, we consider the case $p=2$ and generate samples from three pairs of identical distributions with $n=m=8000$ and $L=n=8000$ projection directions. Repeating this procedure 4000 times produces the histograms shown in Figure 5. The resulting distributions vary significantly across underlying data distributions, which are unknown in practice, making it difficult to compute the critical threshold analytically.

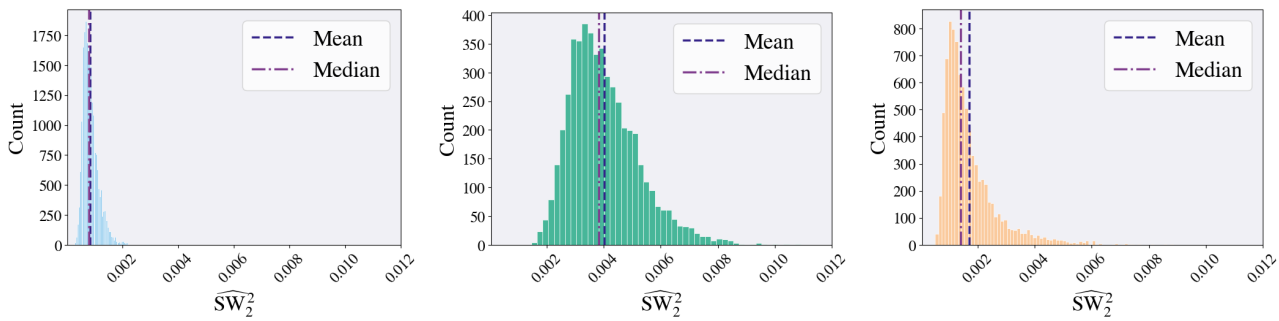


Figure 5: Histograms of the test statistic $\widehat{\mathrm{SW}}_2^2$ computed from 8000 i.i.d. samples drawn respectively from (left) Gaussian $\mathcal{N}(0, I_2)$, (middle) Uniform on $[-1,1]^2$, and (right) a two-component Gaussian mixture $\frac{1}{2}\mathcal{N}(0, I_2) + \frac{1}{2}\mathcal{N}(\mathbf{m}, I_2)$, where $\mathbf{m} = (2.5, 2.5)^\top$.

Recent work by Rodríguez-Vítores et al. (2025) established a Central Limit Theorem (CLT) that characterizes the asymptotic distribution of the empirical Sliced Wasserstein distance. Under suitable regularity conditions, the theorem states that: As $n, m = m(n) \to +\infty$, we obtain that

$$\sqrt{\frac{k\frac{nm}{n+m}}{k+\frac{nm}{n+m}}}\left(\widehat{\mathrm{SW}}_p^p - \mathrm{SW}_p^p(\mu,\nu)\right) \xrightarrow{d} \mathcal{N}\big(0, (1-\tau)w_{\mu,\nu}^2 + \tau\left((1-\lambda)v_{\mu,\nu}^2 + \lambda v_{\nu,\mu}^2\right)\big). \tag{20}$$

Here, $\xrightarrow{d}$ denotes convergence in distribution, $n$ and $m = m(n)$ are the respective sample sizes, and $k = k(n)$ denotes the number of projection directions, and the parameters $\tau$ and $\lambda$ are asymptotic limits defined by

$$\tau = \lim_{n\to+\infty} \frac{k}{k+\frac{nm}{n+m}} \quad \text{and} \quad \lambda = \lim_{n\to+\infty} \frac{n}{n+m}.$$

Moreover, the two components of the asymptotic variance are defined as[2]

$$w_{\mu,\nu}^2 := \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^{2p}\big(\Pi_\#^\theta\mu, \Pi_\#^\theta\nu\big)\, d\sigma(\theta) - \mathrm{SW}_p^{2p}\big(\mu,\nu\big),$$

and

$$v_{\mu,\nu}^2 := \int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} \mathrm{Cov}_\mu\big(\phi^\theta,\, \phi^\eta\big)\, d\sigma(\theta)\, d\sigma(\eta),$$

where $\phi^\theta$ is any c-concave optimal transport potential from $\Pi^\theta_\# \mu$ to $\Pi^\theta_\# \nu$.

As can be observed, the asymptotic variances $w^2_{\mu,\nu}$ and $\nu^2_{\mu,\nu}$ are hard to compute and depend on the underlying unknown distributions. Furthermore, those asymptotic variances are equal to zero under the null hypothesis. As a consequence, it is not possible to construct a meaningful test based on this asymptotic null distribution. More broadly, research on the limiting distribution of the sliced Wasserstein distance remains limited, which makes it difficult to address testing problems through asymptotic analysis.

Finally, the number of observations is often limited due to economical or biological constraints (see, e.g., the discussion about applications in neurosciences in Albert (2015, Section 0.4.1)). Consequently, asymptotic procedures may not be suitable in such small sample size settings. This is one of the reasons why non-asymptotic methods have been developed.

**Non-asymptotic methods.** A classical approach is to determine the threshold of a test using a concentration inequality or a tail bound for the test statistic under the null hypothesis (see, e.g., Gretton et al. (2012, Section 4.2), Wei and Dudley (2012)). This non-asymptotic approach avoids dependence on asymptotic approximations and is therefore less sensitive to distributional assumptions. On the downside, the resulting bounds are usually conservative, leading to Type I errors much smaller than $\alpha$ and lower power. Moreover, many of these guarantees involve constants that are not explicitly determined in theory, making the resulting thresholds difficult to compute in practice (cf. Wang et al. (2021, Section III) and the illustration of Kim et al. (2022, Section 9)).

There exist other data-driven approaches such as the bootstrap (Tibshirani and Efron, 1993) and subsampling (Politis et al., 1999). Many tests based on these methods have been extensively studied (see, e.g., Van Der Vaart and Wellner (1996); Romano (1988); Hu and Lin (2025)). However, as noted in Albert (2015, Section 0.5.2), these methods do not guarantee Type I error control in a non-asymptotic sense, i.e., for small sample sizes — a property that is ensured by permutation-based approaches. As a result, when both methods are applicable and non-asymptotic Type I error control is required, the permutation approach should be preferred.

For the reasons discussed above, we adopt the permutation approach for the two-sample test based on the sliced Wasserstein distance.

# 14 EXISTING APPROACHES FOR CONTROLLING THE RANDOM CRITICAL VALUE OF PERMUTATION-BASED (TWO-SAMPLE) TESTS

In the previous section, we motivated the use of a permutation approach for the sliced Wasserstein distance two-sample test. We now review how the challenges associated with the theoretical analysis of permutation based two-sample tests have been addressed in the literature (see also, for instance, Kim et al. (2022, Section 1.2)).

A central difficulty in analyzing the non-asymptotic power of permutation tests lies in controlling the random critical value of the test, and in particular the dependence structure induced by permutation sampling. To the best of our knowledge, no prior work has addressed this issue in the context of the sliced Wasserstein distance.

In this section, we review two existing approaches for controlling the random rejection threshold in permutation-based tests. Both rely on the structure of U-statistics: the first through a coupling argument (Kim et al., 2022), and the second via a concentration inequality for permuted sums (Albert, 2019). We then explain why current techniques cannot be directly extended to the sliced Wasserstein setting, motivating the development of a new analytical framework—one of the main theoretical contributions of this work.

## 14.1 Coupling technique for U-statistics

Kim et al. (2022, Section 6) investigate the performance of permutation-based procedures with a focus on degenerate second-order U-statistics, a broad class of estimators which encompasses many commonly used two-sample test statistics (Serfling, 2009, Chapter 5). This line of work has inspired several recent theoretical analyses in permutation-based two-sample testing (Schrab et al., 2023; Kim and Schrab, 2023; Choi and Kim, 2024; Chatalic et al., 2025). In order to clarify this approach, we begin by formally introducing second-order U-statistics. We use the same setting and notation as in Section 4.1.

[2]We use the same notation as in the definition of the sliced Wasserstein distance presented in Subsection 2.2

**Definition 29 (Second order U-statistic)** *Let $\mathsf{X}$ be a measurable space. Let $g : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be a measurable symmetric bivariate function, that is, $g(x, y) = g(y, x)$ for all $x, y \in \mathsf{X}$. We introduce*

$$h(y_1, y_2; z_1, z_2) \coloneqq g(y_1, y_2) + g(z_1, z_2) - g(y_1, z_2) - g(y_2, z_1).$$

*Let $\mathbf{i}_n^2$ be the set of all couples drawn without replacement from the set $\{1, \dots, n\}$. Then, the corresponding U-statistic is given by*

$$U_{n,m} \coloneqq \frac{1}{n(n-1)m(m-1)} \sum_{(i_1,i_2)\in \boldsymbol{i}_2^n} \sum_{(j_1,j_2)\in \boldsymbol{i}_2^m} h(Y_{i_1}, Y_{i_2}; Z_{j_1}, Z_{j_2}).$$

*Moreover, given a permutation $\pi \in S_N$ (with $N = n + m$), the permuted U-statistic associated with $\pi$ is defined as*

$$U_{n,m}^{\pi} \coloneqq \frac{1}{n(n-1)m(m-1)} \sum_{(i_1,i_2)\in \boldsymbol{i}_2^n} \sum_{(j_1,j_2)\in \boldsymbol{i}_2^m} h(X_{\pi_{i_1}}, X_{\pi_{i_2}}; X_{\pi_{n+j_1}}, X_{\pi_{n+j_2}}).$$

Assume $n \le m$, and let $L = (l_1, \dots, l_n)$ denote an $n$-tuple drawn uniformly without replacement from the set $\{1, \dots, m\}$. Given $L$, we define the auxiliary test statistic

$$\tilde{U}_{n,m}^{\pi,L} \coloneqq \frac{1}{n(n-1)} \sum_{(k_1,k_2)\in \mathbf{i}_2^n} h\big(X_{\pi_{k_1}}, X_{\pi_{k_2}};\, X_{\pi_{n+l_{k_1}}}, X_{\pi_{n+l_{k_2}}}\big).$$

Note that the statistic $U_{n,m}^{\pi}$ is the conditional expectation of $\tilde{U}_{n,m}^{\pi,L}$ with respect to $L$, given all other random quantities, that is,

$$U_{n,m}^{\pi} = \mathbb{E}_L\left[\tilde{U}_{n,m}^{\pi,L} \,\middle|\, \mathcal{X}_N, \pi\right]. \tag{21}$$

The usefulness of this construction becomes apparent when applying the following lemma from Kim et al. (2022). It is a coupling argument based on a symmetrization trick (Dümbgen, 1998).

**Lemma 30 (Coupling with i.i.d. random variables)** *Let $N \ge 1$, and let $\pi$ be uniformly distributed on the symmetric group $S_N$ of permutations of $\{1, \dots, N\}$.*

*Set $k = \lfloor N/2 \rfloor$, and let $\Delta = (\delta_1, \dots, \delta_k)$ be a random vector of independent Bernoulli variables taking values in $\{0, 1\}^k$. Assume that $\Delta$ is independent of $\pi$. For any permutation $\tau = (\tau_1, \dots, \tau_N) \in S_N$, define $T_\Delta(\tau) \in S_N$ by*

$$\big(T_\Delta(\tau)\big)_{2i-1} \coloneqq \delta_i\, \tau_{2i-1} + (1 - \delta_i)\, \tau_{2i}, \quad \big(T_\Delta(\tau)\big)_{2i} \coloneqq (1 - \delta_i)\, \tau_{2i-1} + \delta_i\, \tau_{2i}, \quad \text{for } i = 1, \dots, k.$$

*Let $\pi' \coloneqq T_\Delta(\pi)$. Then $\pi'$ and $\pi$ are identically distributed.*

For completeness, we provide a proof of this lemma.

**Proof** Assume that $N$ is even. If $N$ is odd, we simply set $(T_\Delta(\tau))_N = \tau_N$ and the proof remains unchanged.

To prove that $\pi'$ and $\pi$ are identically distributed, it suffices to show that, for any arbitrary permutation $\sigma \in S_N$,

$$\mathbb{P}(\pi' = \sigma) = \mathbb{P}\left(\pi = \sigma\right).$$

By applying the law of total probability, we obtain

$$\mathbb{P}\left(\pi' = \sigma\right) = \mathbb{E}\left[\mathbb{P}\left(T_\Delta(\pi) = \sigma \mid \Delta\right)\right].$$

Since $T_\Delta$ is an involution, i.e., $T_\Delta(T_\Delta(\tau)) = \tau$ for all $\tau \in S_N$, the event $\{T_\Delta(\pi) = \sigma\}$ can be equivalently expressed as $\{\pi = T_\Delta(\sigma)\}$. Moreover, since $\pi$ and $\Delta$ are independent, conditioning on $\Delta$ does not affect the distribution of $\pi$. Following these observations, it holds that

$$\mathbb{P}\left(\pi' = \sigma\right) = \mathbb{E}\left[\mathbb{P}\left(\pi = T_\Delta(\sigma) \mid \Delta\right)\right] = \mathbb{E}\left[\mathbb{P}\left(\pi = T_\Delta(\sigma)\right)\right] = \mathbb{E}\left(\frac{1}{N!}\right) = \frac{1}{N!} = \mathbb{P}\left(\pi = \sigma\right).$$

Since this equality holds for every $\sigma \in S_N$, the claim follows. ■

The above lemma implies that the distribution of $\tilde{U}^{\pi,L}_{n,m}$ remains unchanged if we randomly swap $X_{\pi_k}$ and $X_{\pi_{n+l_k}}$ for $k \in \{1, \dots, n\}$. In other words, it allows us to connect the statistic to i.i.d. Bernoulli random variables, which are easier to handle analytically. Moreover, due to the symmetry of $g(x, y)$ and the definition of $h$, this argument naturally extends to i.i.d. Rademacher random variables, viewed as random sign flips.

Indeed, let $\zeta_1, \dots, \zeta_n$ be i.i.d. Rademacher random variables. Then

$$\tilde{U}^{\pi,L,\zeta}_{n,m} := \frac{1}{n(n-1)} \sum_{(k_1,k_2)\in \mathbf{i}^n_2} \zeta_{k_1}\zeta_{k_2}\, h\big(X_{\pi_{k_1}}, X_{\pi_{k_2}};\, X_{\pi_{n+\ell_{k_1}}}, X_{\pi_{n+\ell_{k_2}}}\big) \overset{(d)}{=} \tilde{U}^{\pi,L}_{n,m}. \tag{22}$$

This machinery provides a way to bound the tail probability of $U^{\pi}_{n,m}$. Indeed, for any $\lambda > 0$ and $t > 0$, by applying a Chernoff bound (Vershynin, 2018, Section 2.3) and Jensen's inequality (Vershynin, 2018, Section 1.6) together with (21), we obtain

$$\mathbb{P}_\pi\left(U^{\pi}_{n,m} > \lambda \middle| \mathcal{X}_N\right) \le e^{-\lambda t}\mathbb{E}_\pi\left[\exp\left(\lambda U^{\pi}_{n,m}\right)\middle|\mathcal{X}_N\right] \le e^{-\lambda t}\mathbb{E}_{\pi,L}\left[\exp\left(\lambda \tilde{U}^{\pi,L}_{n,m}\right)\middle|\mathcal{X}_N\right].$$

Then, by (22), it follows that

$$\mathbb{E}_{\pi,L}\left[\exp\left(\lambda \tilde{U}^{\pi,L}_{n,m}\right)\middle|\mathcal{X}_N\right] = \mathbb{E}_{\pi,L,\zeta}\left[\exp\left(\lambda \tilde{U}^{\pi,L,\zeta}_{n,m}\right)\middle|\mathcal{X}_N\right].$$

As a result,

$$\mathbb{P}_\pi\left(U^{\pi}_{n,m} > \lambda \mid \mathcal{X}_N\right) \le \mathbb{E}_{\pi,L,\zeta}\left[\exp\left(\lambda \tilde{U}^{\pi,L,\zeta}_{n,m}\right) \mid \mathcal{X}_N\right].$$

The right-hand side, which involves Rademacher averages, is a well-studied quantity that can be controlled using standard decoupling arguments (Vershynin, 2018, Chapter 6). This, in turn, provides an upper bound on the concentration of $U^{\pi}_{n,m}$ and allows control of the random critical value in the permutation U-statistic test.

Several works on MMD-based two-sample test use this approach by expressing their test statistic as the sum of a U-statistics with respect to the positive definite kernel $g$ associated to the MMD and a controllable remainder term. In contrast, the sliced Wasserstein distance does not relate naturally to U-statistics, preventing a direct application of this approach.

While Kim et al. (2022) rely on the symmetrization trick of Dümbgen (1998) to address the dependence introduced by permutations, Albert (2019) takes a different route and establish a concentration bound for permuted sums, offering an alternative way to control the random quantile term. We describe this second approach in the next section.

### 14.2 Concentration of permuted sums

Continuing our review of methods for controlling the random critical value in permutation-based tests, we now turn to the work of Albert (2019). As shown in Albert et al. (2015, Section 2.2), the test statistic considered in Albert (2019) is a rescaled version of a U-statistic. Although their analysis focuses on independence testing, its contribution extends beyond this specific setting by introducing a distinct framework for analyzing random critical values and highlighting the key role of concentration inequalities in this task.

The main contribution of this work lies in Section 2.2, where the author establishes a concentration inequality for permuted sums in a general setting by leveraging fundamental inequalities for random permutations of Talagrand (1995). In particular, they present the following result

**Lemma 31 (Theorem 2.1 of Albert (2019))** *Let $\{a_{i,j}\}_{1\le i,j\le N}$ be a collection of any real numbers, and $\pi$ be random uniform permutation in $S_N$. Consider $Z_N = \sum_{i=1}^{N} a_{i,\pi(i)}$. Then, for all $x > 0$,*

$$\mathbb{P}\left(|Z_N - \mathbb{E}[Z_N]| \ge 2\sqrt{2\left(\frac{1}{n}\sum_{i,j=1}^{N} a_{i,j}^2\right)} + 2\max_{1\le i,j\le N}\{|a_{i,j}|\}\, x\right) \le 16e^{1/16}\exp\left(-\frac{x}{16}\right). \tag{23}$$

In the final section, the author illustrates the use of this theorem by analyzing the non-asymptotic behavior of a permutation-based independence test introduced in Albert et al. (2015). Before going into the details of how they applied the theorem, we now recall in detail their problem of interest and the corresponding test statistic.

Let $\mathsf{X}$ be a separable space, and let $\mathcal{X}_N = (X_1, \ldots, X_N)$ be i.i.d. samples from a joint distribution $P$ on $\mathsf{X}^2$, where each $X_i = (X_i^1, X_i^2)$ has marginals $P^1$ and $P^2$ corresponding to its coordinates. The goal is to test whether $P = P^1 \otimes P^2$. To this end, they consider the following test statistic, whose motivation is detailed in Albert et al. (2015):

$$T_\delta(\mathcal{X}_N) = \frac{1}{N-1}\left(\sum_{i=1}^{N} \varphi_\delta(X_i^1, X_i^2) \; - \; \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{N} \varphi_\delta(X_i^1, X_j^2)\right).$$

where $\varphi_\delta$ is a measurable real-valued function on $\mathsf{X}^2$ potentially depending on some unknown parameter $\delta$.

Then, for any random permutation $\pi$ uniformly distributed over $\{1, \ldots, N\}$, the corresponding permuted sample is defined as

$$\mathcal{X}_N^\pi = (X_1^\pi, \ldots, X_N^\pi), \qquad \text{where } X_i^\pi = (X_i^1, X_{\pi(i)}^2), \quad \forall\, 1 \le i \le N,$$

and the associated permuted test statistic is given by

$$T_\delta(\mathcal{X}_N^\pi) = \frac{1}{N-1}\left(\sum_{i=1}^{N} \varphi_\delta(X_i^1, X_{\pi(i)}^2) \; - \; \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{N} \varphi_\delta(X_i^1, X_j^2)\right). \tag{24}$$

To apply their concentration bound, they express the permuted test statistic as the difference between a random variable and its expectation. In particular, by introducing $\tilde{Z}(\mathcal{X}_N) = \sum_{i=1}^{N} \varphi_\delta(X_i^1, X_{\pi(i)}^2)$, the permuted test statistic defined in (24) can be rewritten as

$$T_\delta(\mathcal{X}_N^\pi) = \frac{1}{N-1}\Big(\tilde{Z}(\mathcal{X}_N) - \mathbb{E}\left[\tilde{Z}(\mathcal{X}_N) \mid \mathcal{X}_N\right]\Big).$$

This reformulation aligns the left-hand side of the probability inequality in (23) with the permuted test statistic, thereby allowing the authors to directly apply the concentration result to derive an upper bound on its quantile. However, this approach cannot be extended to the case of the sliced Wasserstein distance, since the latter cannot be written as such a centered difference.

From the preceding technical review concerning the control of the conditional quantile of a permuted test statistic, two main insights can be drawn:

- Concentration inequalities are a powerful tool for analyzing conditional quantiles, as they provide sharp non-asymptotic high-probability bounds—typically with exponentially small tails—that lead to the desired logarithmic dependence on the Type I error level $\alpha$, as discussed in Albert (2019, Section 3.2).
- Unlike $U$-statistics, which are centered at zero under permutation, or the statistic in Albert et al. (2015), which can be written as a deviation from its expectation, analyzing the permuted sliced Wasserstein statistic requires explicit control of its expectation under permutation of the samples. We address this challenge in Section 9.

## 15 GAUSSIAN MEAN SHIFT EXPERIMENT

Due to page constraints in the main paper, we present an additional experiment here. In this setting, samples are drawn from $\mu = \mathcal{N}(0, I_{60})$ and $\nu = \mathcal{N}(\mathbf{m}, I_{60})$, where $\mathbf{m} = (0.6, 0.6, 0, \ldots, 0) \in \mathbb{R}^{60}$ is a 60-dimensional vector whose first two entries are equal to 0.6.

Following the experimental setup from Section 6.1, we compare the sliced Wasserstein (SW) tests with the Projected Wasserstein (PW) test (Wang et al., 2021) and the MMD test (Gretton et al., 2012) using linear, Gaussian and Laplace kernels. As observed, the MMD test with a linear kernel shows a significant improvement

in this setting. This improvement can be explained by the fact that for the linear kernel $k(x, y) = x^\top y$, the MMD test reduces to a mean-difference test, making it particularly well-suited for detecting to mean shifts.

It is also observed that, in this scenario, the SW-based tests exhibit lower performance compared to the three MMD-based tests. Nevertheless, the statistical power of SW tests can be considerably improved by increasing the number of projections. This observation motivates the extended investigation presented in Section 6.3.

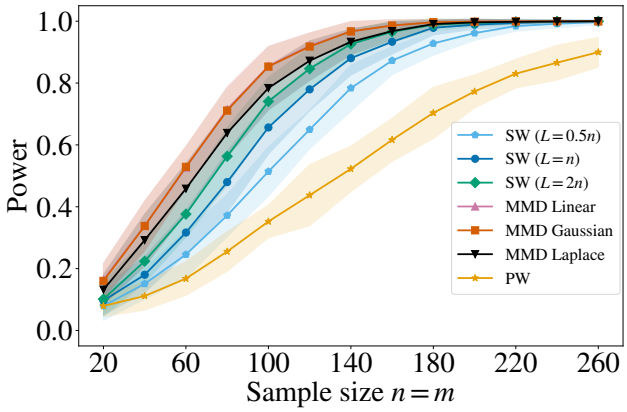


Figure 6: Power vs. number of samples: Gaussian mean shift scenario