Nearly Minimax Optimal Wasserstein Conditional Independence Testing

Matey Neykov[1], Larry Wasserman[*2], Ilmun Kim[*3], and Sivaraman Balakrishnan[*4]

[1,2,4]Department of Statistics & Data Science, Carnegie Mellon University
[3]Department of Statistics and Data Science, Yonsei University

August 21, 2023

Abstract

This paper is concerned with minimax conditional independence testing. In contrast to some previous works on the topic, which use the total variation distance to separate the null from the alternative, here we use the Wasserstein distance. In addition, we impose Wasserstein smoothness conditions which on bounded domains are weaker than the corresponding total variation smoothness imposed, for instance, by Neykov et al. [2021]. This added flexibility expands the distributions which are allowed under the null and the alternative to include distributions which may contain point masses for instance. We characterize the optimal rate of the critical radius of testing up to logarithmic factors. Our test statistic which nearly achieves the optimal critical radius is novel, and can be thought of as a weighted multi-resolution version of the U-statistic studied by Neykov et al. [2021].

1 Introduction

This paper focuses on conditional independence (CI) testing using the Wasserstein distance. CI testing is a fundamental problem in statistics. It has widespread applications in areas such as causal inference and causal discovery [Zhang et al., 2011, Spirtes et al., 2000, Pearl, 2014] and graphical models [Margaritis, 2005, Koller and Friedman, 2009]. In addition it is central to classical statistical concepts such as sufficiency or ancillarity [Dawid, 1979]. On the other hand the Wasserstein distance, and its associated theory of optimal transport, which was originally introduced by Monge [1781], Kantorovich [1942], has recently seen multiple applications in machine learning, and statistical methodology and theory: see for instance [Blanchet and Murthy, 2019] for applications in robust machine learning, [Rubner et al., 2000, Sandler and Lindenbaum, 2011, Li et al., 2013] for applications in image analysis and [Chernozhukov et al., 2017, Hallin et al., 2021, Ghosal and Sen, 2022, Manole et al., 2021] which study optimal transport maps and use them to define multivariate analogues of the quantile of a distribution. Furthermore, also of note are recent uses of optimal transport in nonparametric hypothesis testing problems [Deb and Sen, 2021, Deb et al., 2021], distributional regression [Ghodrati and Panaretos, 2021], generative modeling [Finlay et al., 2020, Onken et al., 2021], fairness in machine learning [Gordaliza

[*]The last three authors are listed randomly.

et al., 2019, Black et al., 2020, De Lara et al., 2021] and statistical applications in the sciences [Komiske et al., 2020].

The Wasserstein distance is flexible and, unlike stronger metrics such as the total variation distance, can be small even when one compares continuous to discrete distributions. This versatility makes it attractive for problems in conditional independence testing where one may not want to assume a priori that the distribution does not contain point masses for example. It is in fact so natural to use the Wasserstein distribution in problems for CI that we are not the first to look into this problem. Warren [2021] develops binning based tests for CI testing problems where the underlying conditional distributions are assumed to be Wasserstein smooth. On the surface, this is similar to what our paper is concerned with: under Wasserstein smoothness assumptions we formulate a binning based statistic. The main difference between our work and Warren [2021] is our goal: we aim to find a (nearly) minimax optimal test statistic and characterize the minimax testing rate, whereas Warren [2021] simply controls the type I and type II errors under certain sufficient conditions. This is a fundamental difference, and our test statistic is markedly distinct from the one used by Warren [2021]: we use a weighted multiresolution U-statistic, whereas Warren [2021] uses a plugin based statistic which compares the Wasserstein distributions on binned samples. This of course makes our analysis quite distinct from that of Warren [2021].

We will now give a high level overview of the minimax approach, inspired by Ingster [1982], Ingster and Suslina [2003], which we undertake. If a null distribution is very close in a certain metric (which in this paper we choose to depend on the Wasserstein distance), to an alternative distribution, tests will have difficulty in distinguishing whether a distribution is coming from the null or the alternative. To remedy this, one can remove distributions which are ε_n-close to the null hypothesis. Our goal is then to discover how small ε_n can be (as a function of the sample size n), so that one can still distinguish the null from the alternative. In addition as we mentioned we impose smoothness conditions both under the null and under the alternative hypothesis. This additional requirement comes as no surprise, since Shah and Peters [2020] proved that under no conditions CI testing is hard in the sense that the power under any alternative distribution of any test that controls the type I error over all (smooth and non-smooth) CI distributions below α is bounded by α.

1.1 Notation

We now summarize commonly used notation throughout the paper.

Definition 1.1 (Total Variation Metric)**.** The total variation (TV) metric between two distributions p, q on a measurable space $(\Omega, \mathcal{F})$ is defined as

$$\mathrm{TV}(p,q) = \sup_{A\in\mathcal{F}} |p(A) - q(A)| = \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\int \left|\frac{dp}{d\nu} - \frac{dq}{d\nu}\right| d\nu,$$

where the last identity assumes ν is a common dominating measure of p and q, i.e., $p \ll \nu$, $q \ll \nu$ and $\frac{dp}{d\nu}, \frac{dq}{d\nu}$ denote the densities of p and q with respect to ν (note here that ν can always be taken as $\nu = p + q$).

We will now formalize our notation for conditional distributions. This notation is the same as the one used in Neykov et al. [2021] but for completeness we provide details here. If the triplet (X, Y, Z) has a distribution $p_{X,Y,Z}$ we will use $p_{X,Y|Z=z}$ to denote the conditional joint distribution of $X, Y|Z = z$. Additionally $p_{X|Z=z}$ and $p_{Y|Z=z}$ will denote the marginal conditional distributions of $X|Z = z$ and $Y|Z = z$ respectively. The marginal distributions will be denoted with p_X, p_Y, p_Z and joint marginal distributions will be denoted with $p_{X,Y}, p_{Y,Z}, p_{X,Z}$. Furthermore, with a slight abuse of notation, $p_{X,Y|Z}(x, y|z)$ and $p_{X|Z}(x|z)$ and $p_{Y|Z}(y|z)$ will denote the densities of these distributions evaluated at the points x, y and z (or the corresponding probability mass functions when X and Y are discrete).

In addition we will use $\lesssim$ and $\gtrsim$ to mean $\leq$ and $\geq$ up to positive universal constants (which may be different from place to place). If both $\lesssim$ and $\gtrsim$ hold we denote this as $\asymp$. For an integer $n \in \mathbb{N}$ we use the convenient shorthand $[n] = \{1, 2, \ldots, n\}$.

Finally, for a real number $r \in \mathbb{R}$ let $\lfloor r \rfloor$ be the largest integer smaller than or equal to r, and let $\lceil r \rceil$ be the smallest integer which is at least r.

1.2 Problem Formulation and Related Works

In this section we formulate the problem precisely and mention some related works. Let $X, Y, Z \in [0,1]^3$ be three random variables. We are interested in testing $H_0 : X \perp\!\!\!\perp Y|Z$ versus the alternative $H_1 : X \not\perp\!\!\!\perp Y|Z$. Define the Wasserstein-1 distance

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|_2 d\gamma(x, y),$$

where Γ denotes the set of all couplings between μ and ν i.e., all joint distributions with marginals μ and ν and $\|\cdot\|_2$ denotes the Euclidean norm. Similarly one can define $W_2(\mu, \nu)$ as

$$W_2(\mu, \nu) = \left[\inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|_2^2 d\gamma(x, y) \right]^{1/2}.$$

We will now state several well-known facts about the W_1, W_2 and TV distances which will be helpful throughout this work.

Fact 1.2. The following statements hold true:

1. For any two probability distributions p, q on $[0,1]^2$:
$$W_2(p, q) \lesssim \mathrm{TV}(p, q).$$
2. For any two probability distributions p, q on $[0,1]^2$:
$$W_1(p, q) \leq W_2(p, q).$$
3. Wasserstein distance is a proper metric, i.e., for three distributions p, q, r on $[0,1]^2$ we have
$$W_i(p, q) \leq W_i(q, r) + W_i(r, p), \quad i \in \{1, 2\}.$$
4. Squared Wasserstein-2 distance is sub-additive on product distributions, i.e., let p_1, p_2, q_1, q_2 be probability distributions on $[0,1]$, then
$$W_2^2(p_1 \times p_2, q_1 \times q_2) \leq W_2^2(p_1, q_1) + W_2^2(p_2, q_2).$$
5. If p, q are probability distributions on $[0,1]^2$ we have
$$W_2^2(p, q) \leq \sqrt{2} W_1(p, q).$$

We defer the proof of this result to the appendix. Let $\mathcal{P}_0$ denote the set of all conditionally independent distributions supported on $[0,1]^3$, i.e. for all $q \in \mathcal{P}_0$: $q_{X,Y|Z} = q_{X|Z} q_{Y|Z}$.

Assumption 1.3. Define the collection of probability distributions

$$\mathcal{P}_0^W(L) := \{p \in \mathcal{P}_0 : W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \leq L|z - z'|, \text{ for all } z, z' \in [0,1]\}.$$

Suppose that under the null hypothesis the distribution belongs to the class $\mathcal{P}_0^W(L)$.

In this paper we work exclusively with W_1 smoothness conditions as in the definition of $\mathcal{P}_0^W(L)$ both under the null, and also under the alternative hypothesis as we will see shortly. Similar smoothness conditions have been used previously to enable binning based approaches to CI testing; see for instance Neykov et al. [2021], Kim et al. [2022b] for total variation smoothness, and also Warren [2021] for Wasserstein smoothness akin to the one we used above. One advantage of the W_1 smoothness in comparison with total variation smoothness as in Neykov et al. [2021], is that on compact domains the W_1 distance is smaller than the total variation up to a constant [See Fact 1.2 (1), and also Lemma 3 and Theorem 6.15 Slawski and Sen, 2022, Villani, 2009, respectively], and therefore, all previous examples suggested in Section 6 of Neykov et al. [2021], which are total variation smooth also satisfy Wasserstein smoothness as defined in Assumption 1.3. Unlike total variation smoothness however, Wasserstein smoothness allows for distributions containing point masses; in other words being a mixture of discrete and continuous distributions may be Wasserstein smooth, while not being total variation smooth as is also pointed out by Warren [2021].

Let $\mathcal{P}_1$ denote the class of all non-conditionally independent distributions i.e., the laws of all random variables $X, Y, Z \in [0,1]^3$ such that $X \not\perp\!\!\!\perp Y|Z$.

Assumption 1.4. Define the collection of alternative distributions $\mathcal{P}_1^W(L,\varepsilon)$ as follows:

$$\mathcal{P}_1^W(L,\varepsilon) := \{p \in \mathcal{P}_1 : W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \leq L|z-z'|, \text{ for all } z, z' \in [0,1], \\ \inf_{q\in\mathcal{P}_0} \mathbb{E}_Z W_2(p_{X,Y|Z}, q_{X,Y|Z}) \geq \varepsilon\}.$$

In the above definition, the expectation over Z is taken with respect to the distribution p_Z which is the Z-marginal of $p_{X,Y,Z}$. We will henceforth assume that the distributions under the alternative hypothesis belong to the class $\mathcal{P}_1^W(L,\varepsilon)$.

We would like to underscore that the W_2 distance is a popular distance which is often considered in practice. For instance Rigollet and Weed [2019] use it to estimate the mean vector in the problem of uncoupled isotonic regression. As the reader can see we are using the W_1 distance to impose smoothness on the distributions while we are using the W_2 distance to impose separation between the null and the alternative. Using distinct measures of smoothness and separation is standard.

See for instance Arias-Castro et al. [2018] where the authors use Hölder smoothness on the densities and L_2 separation in goodness-of-fit problems. Furthermore, since the Wasserstein distance is monotonic (i.e., $W_1 \leq W_2$), the assumed smoothness in W_1 distance is weaker than the respective W_2 smoothness, hence in order to support more distributions we focus on the W_1 smoothness requirement. One final remark that we would like to make on Assumption 1.4 is that for the same amount of separation — ε (up to universal constants) — the Wasserstein separation discards more distributions as compared to the total variation distance. Formally we have

Proposition 1.5. If a distribution p satisfies $\inf_{q\in\mathcal{P}_0} \mathbb{E}_Z W_2(p_{X,Y|Z}, q_{X,Y|Z}) \geq \varepsilon$, then we also have $\inf_{q\in\mathcal{P}_0} \mathrm{TV}(p,q) \gtrsim \varepsilon$.

Proof of Proposition 1.5. To see this first observe that on bounded domains $W_2(p,q) \lesssim \mathrm{TV}(p,q)$ by Fact 1.2 1, where we remind the reader that $\lesssim$ denotes inequality up to absolute constant factors. However, from Lemma B.4 of Neykov et al. [2021] we know

$$\mathrm{TV}(p,q) \geq \mathbb{E}_Z \mathrm{TV}(p_{X,Y|Z}, q_{X,Y|Z})/2 \gtrsim \mathbb{E}_Z W_2(p_{X,Y|Z}, q_{X,Y|Z}).$$

Thus if two distributions p and $q \in \mathcal{P}_0$ satisfy $\mathbb{E}_Z W_2(p_{X,Y|Z}, q_{X,Y|Z}) \geq \varepsilon$, they also satisfy $\mathrm{TV}(p,q) \gtrsim \varepsilon$. □

To summarize, in comparison to Neykov et al. [2021], the Wasserstein separation is "stronger" (by Proposition 1.5) than TV separation, while the Wasserstein smoothness requirement is "weaker" than the corresponding TV smoothness. In order to characterize the complexity of CI testing we use the minimax testing framework, introduced in the work of Ingster and co-authors [Ingster, 1982, Ingster and Suslina, 2003], and which has since then been considered by many authors (see for instance Lepski and Spokoiny [1999], Baraud [2002], Diakonikolas and Kane [2016], Valiant and Valiant [2017], Canonne et al. [2018], Arias-Castro et al. [2018], Canonne [2020], Balakrishnan and Wasserman [2018, 2019], Carpentier and Verzelen [2021], Neykov et al. [2021], Kim et al. [2022a], Albert et al. [2022]). Formally, consider the testing problem

$$H_0 : p \in \mathcal{P}_0^W(L) \text{ vs } H_1 : p \in \mathcal{P}_1^W(L, \varepsilon). \tag{1.1}$$

We define the minimax risk of testing as

$$R_n(\varepsilon) = \inf_{\psi} \left\{ \sup_{p \in \mathcal{P}_0^W(L)} \mathbb{E}_p[\psi(\mathcal{D}_n)] + \sup_{p \in \mathcal{P}_1^W(L,\varepsilon)} \mathbb{E}_p[1 - \psi(\mathcal{D}_n)] \right\}^{1}, \tag{1.2}$$

where the infimum is taken over all Borel measurable test functions $\psi : \text{supp}(\mathcal{D}_n) \mapsto [0, 1]$ (which gives the probability of rejecting the null hypothesis), and $\text{supp}(\mathcal{D}_n)$ is the support of the random variables $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \dots (X_n, Y_n, Z_n)\}$. In the development to follow, we assume that L is a fixed non-zero constant which does not scale with n, and so we do not track the dependence of the critical radius on L.

In the minimax framework our goal is to study the *critical radius* of testing defined as

$$\varepsilon_n(\mathcal{P}_0^W(L), \mathcal{P}_1^W(L, \varepsilon)) = \inf \left\{ \varepsilon : R_n(\varepsilon) \leq \frac{1}{3} \right\}. \tag{1.3}$$

The constant $\frac{1}{3}$ above is arbitrary, and can be chosen as any small constant. The minimax testing radius or the critical radius, corresponds to the smallest radius ε at which there exists *some test* which reliably distinguishes distributions in $\mathcal{H}_0$ from those in $\mathcal{H}_1$ which are appropriately far from $\mathcal{H}_0$. The critical radius provides a fundamental characterization of the statistical difficulty of the hypothesis testing problem in (1.1).

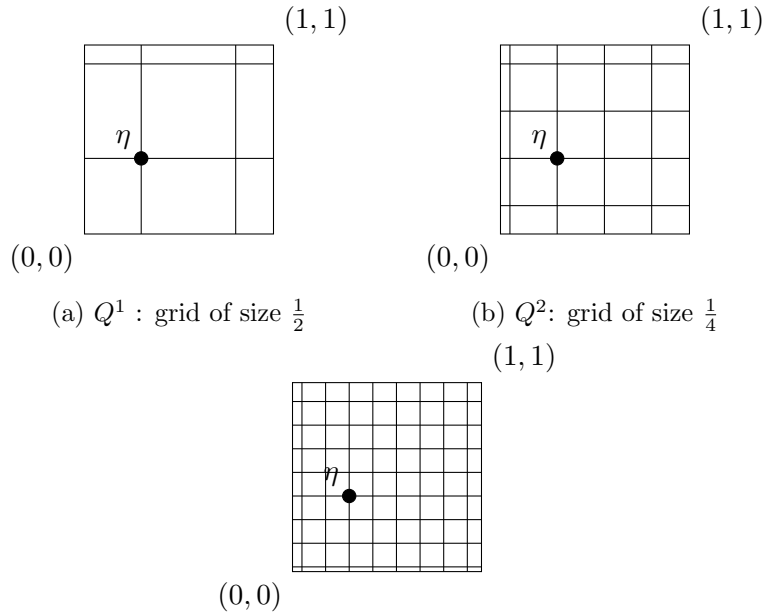
1.3 Organization

The remainder of the paper is structured as follows. In Section 2, we formulate our test and prove it controls the type I and type II errors under an appropriate condition on the radius of separation. In Section 3, we state and prove our main lower bound. Finally, we conclude with a brief discussion of future work in Section 4.

2 Wasserstein Testing

In this section we present the main result of the paper. Our goal is to characterize the critical radius ε_n, defined in (1.3). This involves upper and lower bounding it. Upper bounds are obtained by designing a test and analyzing its Type I and II errors (risk), and lower bounds are obtained via an information theoretic argument. The intuition behind our test construction is rooted in two propositions on the W_2 and W_1 distances given in the papers Weed and Bach [2019], Indyk and Thaper [2003] respectively. These Wasserstein distances can be thought of being approximately weighted "multiresolution" total variation distances (see Lemma 2.1 below).

[1] Here with a slight abuse of notation, we use $\mathbb{E}_p$ to denote expectation under i.i.d. data $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)\}$ where each observation is drawn from p.



(a) Q^1 : grid of size $\frac{1}{2}$

(b) Q^2: grid of size $\frac{1}{4}$

(c) Q^3: grid of size $\frac{1}{8}$

Figure 1: Collection of grids centered at η

Leveraging this result along with tests for distributions which are smooth in total variation [Neykov et al., 2021], we consider a multiresolution test statistic in order to approximates the W_2 separation functional. As we will see, the resulting test yields a nearly (up to logarithmic factors) minimax optimal Wasserstein CI test. The details on the upper bound are given below.

2.1 Upper Bound

Construct $\mathcal{Q}$, a collection of rectangular grids Q^k, $k \in \{1, \ldots, \lceil \log_2(d) \rceil\}$ with side Euclidean length (mostly) $\frac{1}{2^k}$ centered at a fixed point $\eta \in [0,1]^2$. Here d is an integer defined as the number of bins used for the Z variable. Each cell $A_{ij}^k \in Q^k$ is $A_{ij}^k = A_i^k \times A_j^{'k}$ where A_i^k and $A_j^{'k}$ are intervals of size (mostly) $\frac{1}{2^k}$ on $[0,1]$ centered at the projections — η_1 and η_2 — of the point $\eta = (\eta_1, \eta_2)$ on the x and the y axis. We will now formally define the intervals A_i^k for the convenience of the reader. Here the index i ranges in the set $[L]$ where $L = 2 + \lfloor 2^k \eta_1 \rfloor + \lfloor 2^k (1 - \eta_1) \rfloor$. We have

$$
\begin{aligned}
A_1^k &= \left[0,\ \eta_1 - \frac{\lfloor 2^k \eta_1 \rfloor}{2^k}\right) \\
A_i^k &= \left[\eta_1 + \frac{i - 2 - \lfloor 2^k \eta_1 \rfloor}{2^k},\ \eta_1 + \frac{i - 1 - \lfloor 2^k \eta_1 \rfloor}{2^k}\right), \text{ for } i \in \{2, \ldots, L-1\} \\
A_L^k &= \left[\eta_1 + \frac{\lfloor 2^k (1 - \eta_1) \rfloor}{2^k},\ 1\right]
\end{aligned}
$$

Similarly, one can define the interval $A_j^{'k}$ for $j \in [L']$ where $L' = 2 + \lfloor 2^k \eta_2 \rfloor + \lfloor 2^k (1 - \eta_2) \rfloor$. See also Figure 1 for a visualization of three such grids. We now restate and prove a proposition of [Weed and Bach, 2019] adapted to our setting.

Lemma 2.1. For any two distributions p and q on $[0,1]^2$, we have the following inequality:

$$W_2^2(p,q) \lesssim \frac{1}{2^{2\lceil \log_2(d)\rceil+1}} + \sum_{k=1}^{\lceil \log_2(d)\rceil} \frac{1}{2^{2k}} \sum_{A_{ij}^k \in Q^k} |p(A_{ij}^k) - q(A_{ij}^k)|. \tag{2.1}$$

Since the proof of Proposition 2.1 follows directly from the result of Weed and Bach [2019], we defer it to the appendix. We now describe the test used for establishing an upper bound. First draw $N \sim Poi(n/2)$ samples. If $N > n$, accept the null hypothesis. If $N \leq n$, take the first N samples out of the given n samples and discard the rest. We bin the Z support, i.e., $[0,1]$ in d bins of equal size which we denote by $C_1, \ldots, C_d$. This separates the sample $\mathcal{D}_N = \{(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)\}$ into smaller datasets $\mathcal{D}_m = \{(X_i, Y_i) : Z_i \in C_m\}$. Let $|\mathcal{D}_m| = \sigma_m$ denote the sample size of the mth bin. Define the function $g^k((x,y)) = (i,j)$ if and only if $(x,y) \in A_{ij}^k \in Q^k$. Next define the sets $\mathcal{D}_m^k = \{g^k(X_i, Y_i) : Z_i \in C_m\}$ for $m \in [d]$ and $k \in 1, \ldots, \lceil \log_2(d) \rceil$. We now recall the definition of the U-statistic from Neykov et al. [2021]. For two observations i and j and two indices x and y consider the following expression

$$\phi_{ij}(xy) = \mathbb{1}(X_i = x, Y_i = y) - \mathbb{1}(X_i = x)\mathbb{1}(Y_j = y).$$

Note that ϕ takes a value among $\{-1, 0, +1\}$. Next take four observations i, j, k, l and consider the kernel

$$h((X_i,Y_i),(X_j,Y_j),(X_k,Y_k),(X_l,Y_l)) = \frac{1}{4!} \sum_{\pi \in [4!]} \sum_{x,y} \phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy),$$

where π is a permutation of i, j, k, l. Next, construct the corresponding U-statistic

$$U_m(\mathcal{D}_m^k) := \frac{1}{\binom{\sigma_m}{4}} \sum_{i<j<k<l:(i,j,k,l)\in\Sigma_m} h((X_i,Y_i),(X_j,Y_j),(X_k,Y_k),(X_l,Y_l)),$$

where the summation is over choosing 4 distinct elements from Σ_m, where Σ_m denotes the set of distinct indices in the set $\mathcal{D}_m$. We now define the test statistic:

$$T := \mathbb{E}_\eta \sum_{k=1}^{\lceil \log_2(d)\rceil} \frac{1}{2^{2k}} \sum_{m\in[d]} U(\mathcal{D}_m^k)\mathbb{1}(\sigma_m \geq 4)\sigma_m,$$

where the expectation above over η is taken with respect to uniformly sampling η on the grid points of a square grid of side Euclidean length equal to $1/(2^{\lceil \log_2(d)\rceil+1})$ on $[0,1]^2$ centered at $\mathbf{0} = (0,0)$. We then define the test

$$\psi_\tau(\mathcal{D}_N) = \mathbb{1}(T \geq \tau)\mathbb{1}(N \leq n). \tag{2.2}$$

Remark 2.2 (On computing the test $\psi_\tau(\mathcal{D}_N)$)**.** According to a careful analysis in Section 3.3 of Kim et al. [2023] calculating $U(\mathcal{D}_m^k)$ can be done in $O(\sigma_m)$ operations. This implies that (for a fixed η) calculating $\sum_{m\in[d]} U(\mathcal{D}_m^k)\mathbb{1}(\sigma_m \geq 4)\sigma_m$ takes at most $O(n)$ time; then $\sum_{k=1}^{\lceil \log_2(d)\rceil} \frac{1}{2^{2k}} \sum_{m\in[d]} U(\mathcal{D}_m^k)\mathbb{1}(\sigma_m \geq 4)\sigma_m$ takes $O(\log_2(d)n)$ time. Since in the end we set $d \asymp n^{2/5}$ for a fixed η we have $O(n\log_2(n))$ operations. Finally, since η belongs to a grid of at most $16d^2$ points, we have that the computational cost is $O(d^2\log_2(d)n) = O(n^{9/5}\log_2(n))$. This is bigger than linear complexity so it can be prohibitive for a large n. However we note that the computational complexity of calculating T is better than quadratic time.

We are now ready to state the main result of the paper.

Theorem 2.3. Take $d = \lceil n^{2/5} \rceil$ and set $\tau = \zeta\sqrt{d}\log_2^2 d$ for a sufficiently large constant ζ. Suppose that $\varepsilon > \frac{c(\log_2 d)^{3/4}}{d^{1/2}}$, for a sufficiently large constant c. Then

$$\sup_{p \in \mathcal{P}_0^W(L)} \mathbb{E}_p \psi_\tau(\mathcal{D}_N) \leq \frac{1}{10},$$
$$\sup_{p \in \mathcal{P}_1^W(L,\varepsilon)} \mathbb{E}_p(1 - \psi_\tau(\mathcal{D}_N)) \leq \frac{1}{10} + \exp(-n/8).$$

Setting $d = \lceil n^{2/5} \rceil$, the above result establishes an upper bound for the critical radius as

$$\varepsilon_n(\mathcal{P}_0^W(L), \mathcal{P}_1^W(L,\varepsilon)) \leq c_1 \frac{(\log_2 n)^{3/4}}{n^{1/5}},$$

where c_1 is some positive constant. We now compare this rate to the rates given in Neykov et al. [2021]. There are two main results in Neykov et al. [2021] regarding the separation radius.

1. First we comment on the "fully" continuous setting. In this setting, Neykov et al. [2021] assume that the distributions are TV smooth, i.e. that the conditional distributions are Lipschitz in the TV sense as a function of the conditioning variable Z. Additionally, they assume that the distributions $X, Y|Z = z$ have Hölder continuous density functions with exponent s for all z (see Definitions 2.3 and 2.4 in Neykov et al. [2021] for more details). The critical radius in their work scales as $n^{-2s/(5s+2)}$, which is faster than the $n^{-1/5}$ rate we obtain in this paper, for all sufficiently large s values. In this paper we assume that the conditional distributions are Lipschitz in the W_1 sense, but in stark contrast to the work of Neykov et al. [2021] we do not require additional smoothness on the distributions (such as Hölder smoothness). While this results in a slower rate, we earn flexibility in terms of the allowed distributions. Indeed, this flexibility is the main benefit afforded by testing using the Wasserstein distance. When testing under separation in the TV metric, even problems simpler than CI testing, such as goodness-of-fit testing, are impossible without additional smoothness assumptions [Balakrishnan and Wasserman, 2019]. This is however not the case for testing with separation in the Wasserstein distance [Ba et al., 2011], which is a tractable task even without smoothness assumptions.
2. The Wasserstein smoothness assumptions we impose can also support discrete distributions, and hence it is also sensible to compare our rates with the discrete case considered by Neykov et al. [2021]. The rate in the TV smoothness setting is $n^{-2/5}$, which is faster than the $n^{-1/5}$ that we established above. We can conclude that even with stronger separation requirement we impose, the problem of Wasserstein testing is harder than TV testing in the discrete case considered by Neykov et al. [2021].

The remaining of this section is devoted to the proof of Theorem 2.3.

2.2 Proof of Theorem 2.3

Similarly to the proof of Theorem 5.2 of Neykov et al. [2021], it suffices to show the result assuming that $N \sim Poi(n)$. We will analyze the expectation and variance of T in Section 2.2.1 and Section 2.2.3, respectively. We do so in order to apply Chebyshev's inequality and control the risk from above (see Section 2.3).

2.2.1 Analysis of the Expectation

In this section we are concerned with controlling the expectation $\mathbb{E}T$ from below and above under the alternative and the null hypothesis respectively. Fix $k \in \{1, \ldots, \lceil \log_2(d) \rceil\}$. Starting

with the expectation, conditional on σ_m with $\sigma_m \geq 4$, we have that

$$\mathbb{E}[U(\mathcal{D}_m^k)|\sigma_m] = \sum_{i,j}(q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m))^2,$$

where $q_{ij}^k(m) = P_{X,Y|Z\in C_m}(A_{ij}^k|Z \in C_m)$ and $q_{i\cdot}^k(m) = \sum_j q_{ij}^k(m) = P_{X|Z\in C_m}(A_i^k|Z \in C_m)$, and similarly for $q_{\cdot j}^k(m)$. With a slight abuse of notation, we define the expression $U(\mathcal{D}_m^k) := \mathbb{E}[U(\mathcal{D}_m^k)|\sigma_m] := \sum_{i,j}(q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m))^2$ even when $\sigma_m < 4$ even though in this case the U-statistic $U(\mathcal{D}_m^k)$ is not well defined. This is a legitimate operation, since our test statistic T does not "see" the values of the U-statistic for m such that $\sigma_m < 4$. In other words, since the indicator $\mathbb{1}(\sigma_m \geq 4)U(\mathcal{D}_m^k) = 0$ when $\sigma_m < 4$ we can define the value of $U(\mathcal{D}_m^k)$ to be $\sum_{i,j}(q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m))^2$. Let $p_m = \mathbb{P}(Z \in C_m)$.

Analysis under the Alternative Hypothesis. The goal of this section is to lower bound $\mathbb{E}T$ under the alternative. We start by looking into the following expression

$$\begin{aligned}
\sum_{m\in[d]} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]}p_m &= \sum_{m\in[d]} \sqrt{\mathbb{E}_\eta \sum_{i,j}\left(q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)\right)^2}p_m \\
&\geq \mathbb{E}_\eta \sum_{m\in[d]} \sqrt{\sum_{i,j}\left(q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)\right)^2}p_m \\
&\geq \sum_{m\in[d]} \frac{\sum_{i,j}\mathbb{E}_\eta|q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)|}{2^k+1}p_m \\
&\geq \sum_{m\in[d]} \frac{\sum_{i,j}\mathbb{E}_\eta|q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)|}{2^{k+1}}p_m \qquad (2.3)
\end{aligned}$$

where we used Jensen's inequality, the fact that $\sqrt{\sum_{i=1}^l a_i^2} \geq \sum_{i=1}^l a_i/\sqrt{l}$ for any real numbers $a_i \in \mathbb{R}$, and the fact that there are at most $(2^k+1)^2$ cells in Q^k (here observe that $L \leq 2^k+2$ (as defined in the beginning of Section 2.1); however, L can be 2^k+2 only when $\lfloor 2^k\eta_1 \rfloor$ and $\lfloor 2^k(1-\eta_1)\rfloor$ are both integers in which case A_1^k, A_L^k are $\varnothing$ so that we effectively have 2^k intervals in that case; hence we have at most $(2^k+1)^2$ cells in Q^k since the same logic is valid for L'). We will now need the following result which quantifies the error in approximation of the expected W_2 with its binned counterpart.

Lemma 2.4. If the distribution $p_{X,Y,Z}$ is Wasserstein 1-smooth, i.e., $W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \leq L|z-z'|$ we have that

$$\begin{aligned}
\varepsilon &\leq \inf_{q\in\mathcal{P}_0} \mathbb{E}_Z W_2(p_{X,Y|Z}, q_{X,Y|Z}) \leq \int W_2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z})dP(z) \\
&\leq \sum_{m\in[d]} W_2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m}p_{Y|Z\in C_m})p_m + \kappa(L\max_{m\in[d]}\operatorname{diam}(C_m))^{1/2},
\end{aligned}$$

where κ is an absolute constant.

Remark 2.5. By the elementary inequality $(a+b)^2 \leq 2a^2+2b^2$, and the convexity of $x \mapsto x^2$

we have

$$\begin{aligned}\varepsilon^2 &\le \Big(\inf_{q\in\mathcal{P}_0} \mathbb{E}_Z W_2(p_{X,Y|Z}, q_{X,Y|Z})\Big)^2 \\ &\le \Big(\sum_{m\in[d]} W_2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m} p_{Y|Z\in C_m}) p_m + \kappa L^{1/2} \max_{m\in[d]} \operatorname{diam}(C_m)^{1/2}\Big)^2 \\ &\lesssim \sum_{m\in[d]} W_2^2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m} p_{Y|Z\in C_m}) p_m + \kappa^2 L \max_{m\in[d]} \operatorname{diam}(C_m). \end{aligned} \tag{2.4}$$

We defer the proof of Lemma 2.4 to the Appendix. Continuing the bound (2.3) we conclude that

$$\begin{aligned}\sum_{k=1}^{\log_2(d)} \sum_{m\in[d]} \frac{1}{2^k} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} p_m &\ge \sum_k \sum_{m\in[d]} \frac{\sum_{i,j} \mathbb{E}_\eta |q_{ij}^k(m) - q_{i\cdot}^k(m) q_{\cdot j}^k(m)|}{2^{2k+1}} p_m \\ &\ge C \sum_{m\in[d]} W_2^2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m} p_{Y|Z\in C_m}) p_m - \frac{1}{d^2} \\ &\ge C\varepsilon^2 - \frac{C'L}{d} - \frac{1}{d^2} =: \Upsilon, \end{aligned} \tag{2.5}$$

where C is some absolute constant from (2.1), and the $-\frac{1}{d^2}$ comes from the term $\frac{1}{2^{2\lceil \log_2(d)\rceil}}$ where the term $\frac{C'L}{d}$ comes from Lemma 2.4, and more specifically from the last term on the right hand side of (2.4). Note also that the inequality of Lemma 2.1 holds for any η, which means that it also holds in expectation.

Next by Lemma 3.1 of Canonne et al. [2018] we have

$$\begin{aligned}\sum_k \frac{1}{2^{2k}} \sum_{m\in[d]} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \ge 4)] &\ge \gamma \sum_k \frac{1}{2^{2k}} \sum_{m:(np_m)>1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] np_m \\ &\quad + \gamma \sum_k \frac{1}{2^{2k}} \sum_{m:(np_m)\le 1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] (np_m)^4, \end{aligned}$$

for an absolute constant γ. Since by (2.5) we have that $\sum_{k=1}^{\log_2(d)} \sum_{m\in[d]} \frac{1}{2^k} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} p_m \ge \Upsilon$ we have that either

$$\sum_{k=1}^{\log_2(d)} \frac{1}{2^k} \sum_{m:(np_m)>1} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m \ge \frac{n\Upsilon}{2}, \text{ or} \tag{2.6}$$

$$\sum_{k=1}^{\log_2(d)} \frac{1}{2^k} \sum_{m:(np_m)\le 1} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m \ge \frac{n\Upsilon}{2}. \tag{2.7}$$

We now consider two cases:

i. In the first case we assume (2.6) (where we remind the reader that Υ is defined in (2.5)). By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\sum_{m:(np_m)>1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] np_m &\ge \frac{\big(\sum_{m:(np_m)>1} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m\big)^2}{\sum_{m:(np_m)>1} np_m} \\ &\ge \frac{\big(\sum_{m:(np_m)>1} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m\big)^2}{n}. \end{aligned}$$

Hence

$$\sum_k \frac{1}{2^{2k}} \sum_{m:(np_m)>1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] np_m \geq \sum_k \frac{(\sum_{m:(np_m)>1} \frac{1}{2^k} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m)^2}{n}$$
$$\geq \frac{(\sum_k \sum_{m:(np_m)>1} \frac{1}{2^k} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m)^2}{n\lceil \log_2 d \rceil}$$
$$\gtrsim \frac{n\Upsilon^2}{\lceil \log_2 d \rceil}.$$

ii. In the second case we suppose (2.7) holds. Note that for any non-negative sequences $\{a_m\}_{m=1}^n$ and $\{b_m\}_{m=1}^n$, Jensen's inequality yields

$$\sum_{m=1}^n \frac{a_m^{1/3}}{\sum_{j=1}^n a_j^{1/3}} a_m^{2/3} b_m^4 \geq \Big(\sum_{m=1}^n \frac{a_m^{1/3}}{\sum_{j=1}^n a_j^{1/3}} a_m^{1/6} b_m \Big)^4.$$

Taking

$$a_m = \sum_k \frac{1}{2^{2k}} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] \quad \text{and} \quad b_m = np_m,$$

we have

$$\Big(\sum_{m:(np_m)\leq 1} \Big(\sum_k \frac{1}{2^{2k}} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] \Big)^{1/3} \Big)^3 \sum_k \frac{1}{2^{2k}} \sum_{m:(np_m)\leq 1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] (np_m)^4$$
$$\geq \Big(\sum_{m:(np_m)\leq 1} \sqrt{\sum_k \frac{1}{2^{2k}} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m \Big)^4,$$

and therefore

$$\sum_k \frac{1}{2^{2k}} \sum_{m:(np_m)\leq 1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] (np_m)^4 \gtrsim \Big(\sum_{m:(np_m)\leq 1} \sqrt{\sum_k \frac{1}{2^{2k}} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} np_m \Big)^4 / d^3,$$

since $\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] \leq \mathbb{E}_\eta\big(\sum_{x,y} |q_{xy}(m) - q_{x\cdot}(m) q_{\cdot y}(m)| \big)^2 \leq 4$, and the summation over k reduces to a converging geometric series and finally $|\{m : (np_m) \leq 1\}| \leq d$. Now, by the Cauchy–Schwarz inequality,

$$\sqrt{\sum_k \frac{1}{2^{2k}} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]} \geq \sum_k \frac{1}{2^k} \sqrt{\mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m])} / \lceil \log_2 d \rceil^{1/2}$$

so we conclude

$$\sum_k \frac{1}{2^{2k}} \sum_{m:(np_m)\leq 1} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m] (np_m)^4 \gtrsim \frac{(n\Upsilon)^4}{d^3 \lceil \log_2 d \rceil^2}.$$

Combining the above results, we have established that under the alternative,

$$\mathbb{E}[T] \gtrsim \min\Big\{ \frac{n\Upsilon^2}{\lceil \log_2 d \rceil}, \frac{(n\Upsilon)^4}{d^3 \lceil \log_2 d \rceil^2} \Big\}. \tag{2.8}$$

Analysis under the Null Hypothesis. Next we will upper bound the expectation of T under the null hypothesis:

$$\begin{aligned}\sum_{m\in[d]} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]\mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] &\leq n \sum_{m\in[d]} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]p_m. \\ &= n \sum_{m\in[d]} \mathbb{E}_\eta \sum_{i,j} (q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m))^2 p_m \\ &\leq n \sum_{m\in[d]} \mathbb{E}_\eta (\sum_{i,j} |q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)|)^2 p_m \\ &\lesssim n \sum_{m\in[d]} \mathbb{E}_\eta (\sum_{i,j} |q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)|) p_m,\end{aligned}$$

where we remind the reader that we assume $\mathbb{E}[U(\mathcal{D}_m^k)|\sigma_m] = (q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m))^2$ for all m (even though the value of $U(\mathcal{D}_m^k)$ is technically only defined for $m : \sigma_m \geq 4$). We now remind the reader that $q_{ij}^k(m) = P_{X,Y|Z\in C_m}(A_{ij}^k|Z \in C_m)$ and $q_{i\cdot}^k(m) = \sum_j q_{ij}^k(m) = P_{X|Z\in C_m}(A_i^k|Z \in C_m)$, and similarly for $q_{\cdot j}^k(m)$. Next we will handle the expression

$$\begin{aligned}&\sum_{i,j} |q_{ij}^k(m) - q_{i\cdot}^k(m)q_{\cdot j}^k(m)| \\ &\quad= \sum_{i,j} \left| \int_{C_m} P(A_{ij}^k|Z=z)d\widetilde{P}(z) - \int_{C_m} P(A_i^k|Z=z)d\widetilde{P}(z) \int_{C_m} P(A_j^{'k}|Z=z)d\widetilde{P}(z) \right| \\ &\quad= \sum_{i,j} \left| \int_{C_m} P(A_i^k|Z=z)P(A_j^{'k}|Z=z)d\widetilde{P}(z) - \int_{C_m} P(A_i^k|Z=z)d\widetilde{P}(z) \int_{C_m} P(A_j^{'k}|Z=z)d\widetilde{P}(z) \right| \\ &\quad\leq \int_{C_m} \sum_i \left| P_{X|Z=z}(A_i^k) - \int_{C_m} P_{X|Z=z}(A_i^k)d\widetilde{P}(z) \right| \\ &\quad\times \sum_j \left| P_{Y|Z=z}(A_j^{'k}) - \int_{C_m} P_{X|Z=z}(A_j^{'k})d\widetilde{P}(z) \right| d\widetilde{P}(z),\end{aligned}$$

by Jensen's inequality and where $\widetilde{P}(z) = dP(z)/P(Z \in C_m)$.

We will now argue that the above is smaller than or equal to the product of total variations. Take the first term. By Jensen's inequality

$$\begin{aligned}&\sum_i \left| P_{X|Z=z}(A_i^k) - \int_{C_m} P_{X|Z=z}(A_i^k)d\widetilde{P}(z) \right| \leq \int_{C_m} \sum_i \left| P_{X|Z=z}(A_i^k) - P_{X|Z=z'}(A_i^k) \right| d\widetilde{P}(z') \\ &= 2 \int_{C_m} d_{\mathrm{TV}}(P_{X|Z=z}^k, P_{X|Z=z'}^k)d\widetilde{P}(z'),\end{aligned}$$

where P^k denotes the discretized distributions on the grid. We now have

$$
\begin{aligned}
&\mathbb{E}_\eta \sum_k \frac{1}{2^{2k}} n \sum_{m\in[d]} \left(\sum_{i,j} |q^k_{ij}(m) - q^k_{i\cdot}(m) q^k_{\cdot j}(m)| \right) p_m \\
&\quad \le \sum_{m\in[d]} \mathbb{E}_\eta \sum_k \frac{1}{2^{2k}} n \int_{C_m}\int_{C_m}\int_{C_m} 4 d_{\mathrm{TV}}(P^k_{X|z}, P^k_{X|z'}) d_{\mathrm{TV}}(P^k_{Y|z}, P^k_{Y|z''}) d\widetilde{P}(z') d\widetilde{P}(z'') d\widetilde{P}(z) p_m \\
&\quad \le 4n \sum_{m\in[d]} \int_{C_m}\int_{C_m}\int_{C_m} \left(\mathbb{E}_{\eta_1} \sum_k \frac{1}{2^k} d_{\mathrm{TV}}(P^k_{X|z}, P^k_{X|z'}) \right) \times \\
&\qquad\qquad \left(\mathbb{E}_{\eta_2} \sum_k \frac{1}{2^k} d_{\mathrm{TV}}(P^k_{Y|z}, P^k_{Y|z''}) \right) d\widetilde{P}(z') d\widetilde{P}(z'') d\widetilde{P}(z) p_m \\
&\quad \lesssim n(\log_2 d)^2/d^2,
\end{aligned}
$$

since by Lemma 2.6, proved below, the summations are bounded as:

$$
\begin{aligned}
2\mathbb{E}_{\eta_1} \sum_k 1/2^k d_{\mathrm{TV}}(P^k_{X|z}, P^k_{X|z'}) &\le (\lceil \log_2(d) \rceil + 1) 4 \left(W_1(P_{X|z}, P_{X|z'}) + \frac{1}{2^{\lceil \log_2(d) \rceil + 1}} \right) \\
&\lesssim \log_2 d (L/d + 1/d),
\end{aligned}
$$

using the Wasserstein smoothness as in Assumption 1.3 and also inequality (A.2). Hence under the null, we have

$$
\mathbb{E}[T] \le \frac{C(\log_2 d)^2 n}{d^2}.
$$

2.2.2 On a Lemma of Indyk and Thaper [2003]

We now prove a modified result of Indyk and Thaper [2003]. The main added twist is the fact that η need not be uniform on $[0,1]^q, q \in \mathbb{N}$ but can be in fact taken to be uniformly distributed on a sufficiently small grid. This has an important practical implication as it enables calculating our test statistic. Although our result contains this additional complication, the proof still follows the idea of Indyk and Thaper [2003]. Let

$$
\mathcal{Q} = \left\{ Q^k : k \in 0, 1, \ldots, \lceil \log_2(\frac{1}{\varphi}) \rceil \right\},
$$

be a collection of grids on $[0,1]^q$ for $q \in \mathbb{N}$, with side (Euclidean) length $\frac{1}{2^k}$, centered at the point $\eta \in [0,1]^q$ (we will only use the result when $q = 1$). Here η lies on a grid of side length $\frac{1}{2^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1}}$ centered at 0. Let $\mathbb{E}_\eta$ denote the expectation with respect to η uniformly sampled on the aforementioned grid.

Lemma 2.6. Then we have

$$
\mathbb{E}_\eta \sum_{k\in[|\mathcal{Q}|]} \frac{1}{2^k} \sum_{C\in Q^k} |\mu(C) - \nu(C)| \le (\lceil \log_2(\frac{1}{\varphi}) \rceil + 1) 4q \left(W_1(\mu,\nu)/\sqrt{q} + \frac{1}{2^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1}} \right).
$$

Proof. Define

$$
S_k = \left\{ (x,y) : \frac{\sqrt{q}}{2^{k+1}} < \|x - y\|_2 \le \frac{\sqrt{q}}{2^k} \right\},
$$

for $k = 0, 1, \ldots, \lceil \log_2(\frac{1}{\varphi}) \rceil$, and let $S_{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1} = \left\{ (x,y) : \|x-y\|_2 \leq \frac{\sqrt{q}}{2^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1}} \right\}$. Let γ be an optimal coupling for the Wasserstein distance $W_1(\mu,\nu)$. By definition, we have the following bound

$$W_1(\mu,\nu) \geq \sum_{\ell=0}^{\lceil \log_2(\frac{1}{\varphi}) \rceil} \frac{\sqrt{q}}{2^{\ell+1}} \int_{S_\ell} d\gamma(x,y) = \sum_{\ell=0}^{\lceil \log_2(\frac{1}{\varphi}) \rceil} \frac{\sqrt{q}}{2^{\ell+1}} \gamma(S_\ell). \tag{2.9}$$

We will now re-express the multi-resolution L_1 distance in terms of γ. We have

$$\begin{aligned}
\mathbb{E}_\eta \sum_{k \in [|\mathcal{Q}|]} \frac{1}{2^k} \sum_{C \in Q^k} |\mu(C) - \nu(C)| &= \sum_{k \in [|\mathcal{Q}|]} \frac{1}{2^k} \mathbb{E}_\eta \sum_{C \in Q^k} \left| \int_{C \times [0,1]^q} d\gamma(x,y) - \int_{[0,1]^q \times C} d\gamma(x,y) \right| \\
&= \sum_{k \in [|\mathcal{Q}|]} \frac{1}{2^k} \mathbb{E}_\eta \sum_{C \in Q^k} \left| \int_{C \times C^c} d\gamma(x,y) - \int_{C^c \times C} d\gamma(x,y) \right| \\
&\leq \sum_{k \in [|\mathcal{Q}|]} \frac{1}{2^k} \mathbb{E}_\eta \sum_{C \in Q^k} (\gamma(C \times C^c) + \gamma(C^c \times C)) \\
&= \sum_{k \in [|\mathcal{Q}|]} \frac{1}{2^{k-1}} \mathbb{E}_\eta \sum_{C \in Q^k} \gamma(C \times C^c),
\end{aligned} \tag{2.10}$$

where the last identity follows since each two distinct sets $C_1, C_2 \in Q^k$ we have $\gamma(C_1 \times C_2)$ and $\gamma(C_2 \times C_1)$ appearing once in each of the two summations. Next we will control the expression

$$\begin{aligned}
\mathbb{E}_\eta \sum_{C \in Q^k} \gamma(C \times C^c) &= \mathbb{E}_\eta \sum_{C \in Q^k} \sum_{\ell=0}^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1} \gamma(C \times C^c \cap S_\ell) \\
&= \mathbb{E}_\eta \sum_{C \in Q^k} \sum_{\ell=0}^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1} \gamma(S_\ell) \int \mathbb{1}_{C \times C^c}(x,y) d\gamma(x,y|S_\ell) \\
&= \sum_{\ell=0}^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1} \gamma(S_\ell) \int \mathbb{E}_\eta \sum_{C \in Q^k} \mathbb{1}((x,y) \in C \times C^c) d\gamma(x,y|S_\ell).
\end{aligned}$$

Note that $\mathbb{E}_\eta \sum_{C \in Q^k} \mathbb{1}((x,y) \in C \times C^c) = \mathbb{P}_\eta((x,y) \in \cup_{C \in Q^k} C \times C^c)$ is the probability that the edge $(x,y) \in S_\ell$ "crosses" the grid Q^k. Let $z_1, \ldots, z_q$ be the lengths of the Euclidean projections of the vector $y - x$ on the axis. The grid is crossed if and only if any of the projections crosses a side of the grid. By the union bound this happens with probability at most

$$\begin{aligned}
\sum_{i \in [q]} \frac{z_i + s}{\frac{1}{2^k}} &= 2^k \sum_{i \in [d]} z_i + 2^k q s \leq 2^k \sqrt{q} \sqrt{\sum_{i \in [d]} z_i^2} + 2^k q s \\
&= 2^k \sqrt{q} \|x - y\|_2 + 2^k q s \leq \frac{2^k q}{2^\ell} + 2^k q s,
\end{aligned}$$

where the last bound holds since $(x,y) \in S_\ell$ and $s = \frac{1}{2^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1}}$ is the length of the grid for η. We conclude that

$$\sum_{C \in Q^k} \mathbb{E}_\eta \gamma(C \times C^c) \leq 2^k q s + \sum_{\ell=0}^{\lceil \log_2(\frac{1}{\varphi}) \rceil + 1} \gamma(S_\ell) \frac{2^k q}{2^\ell}.$$

Going back to (2.10) we have

$$\begin{aligned}\mathbb{E}_\eta \sum_{k\in[|\mathcal{Q}|]} \frac{1}{2^k} \sum_{C\in\mathcal{Q}^k} |\mu(C)-\nu(C)| &\le \sum_{k\in[|\mathcal{Q}|]} \frac{1}{2^{k-1}}\Big(2^k qs + \sum_{\ell=0}^{\lceil\log_2(\frac{1}{\varphi})\rceil+1} \gamma(S_\ell)\frac{2^k q}{2^\ell}\Big)\\ &\le (\lceil\log_2(\frac{1}{\varphi})\rceil+1)2qs\\ &+ \sum_{k\in[|\mathcal{Q}|]} \sum_{\ell=0}^{\lceil\log_2(\frac{1}{\varphi})\rceil} \gamma(S_\ell)\frac{4d}{2^{\ell+1}} + \sum_{k\in[|\mathcal{Q}|]} \gamma(S_{\lceil\log_2(\frac{1}{\varphi})\rceil+1})\frac{2d}{2^{\lceil\log_2(\frac{1}{\varphi})\rceil+1}}\\ &\le (\lceil\log_2(\frac{1}{\varphi})\rceil+1)4q\Big(s/2 + \frac{W_1(\mu,\nu)}{\sqrt{q}} + \frac{1}{2^{\lceil\log_2(\frac{1}{\varphi})\rceil+2}}\Big),\end{aligned}$$

where we used (2.9) in the above inequality. Recalling that $s = \frac{1}{2^{\lceil\log_2(\frac{1}{\varphi})\rceil+1}}$, the above can be made smaller than

$$(\lceil\log_2(\frac{1}{\varphi})\rceil+1)4q\Big(\frac{W_1(\mu,\nu)}{\sqrt{q}} + \frac{1}{2^{\lceil\log_2(\frac{1}{\varphi})\rceil+1}}\Big),$$

as claimed. □

2.2.3 Analysis of the Variance

We now turn to the analysis of the variance of the test statistic T. First of all, the rule of total variance ensures that

$$\operatorname{Var} T = \mathbb{E}[\operatorname{Var}[T|\sigma]] + \operatorname{Var}[\mathbb{E}[T|\sigma]],$$

where $\sigma = (\sigma_m)_{m\in[d]}$. The first term is

$$\operatorname{Var}[T|\sigma] = \sum_{m,j\in[d]} \operatorname{Cov}(T^m, T^j|\sigma_m,\sigma_j) = \sum_{m\in[d]} \operatorname{Var}(T^m|\sigma_m),$$

where $T^m = \sum_k 1/2^{2k}\mathbb{E}_\eta U(\mathcal{D}_m^k)\mathbb{1}(\sigma_m \ge 4)\sigma_m$. Since $\operatorname{Var}(\sum_{i\in[k]} X_i) \le k\sum \operatorname{Var}(X_i)$ (which follows by the fact that $\operatorname{Cov}(X,Y) \le \operatorname{Var}(X)/2 + \operatorname{Var}(Y)/2$) we have

$$\begin{aligned}\operatorname{Var}(T^m|\sigma_m) &\le (\log_2 d)\sum_k \operatorname{Var}(1/2^{2k}\mathbb{E}_\eta U(\mathcal{D}_m^k)\mathbb{1}(\sigma_m\ge 4)\sigma_m|\sigma_m)\\ &= (\log_2 d)\sum_k \mathbb{1}(\sigma_m\ge 4)\sigma_m^2 \operatorname{Var}(1/2^{2k}\mathbb{E}_\eta U(\mathcal{D}_m^k)|\sigma_m).\end{aligned}$$

Now, $\operatorname{Var}(\mathbb{E}_\eta U(\mathcal{D}_m^k)|\sigma_m) \le \mathbb{E}_\eta \operatorname{Var}(U(\mathcal{D}_m^k)|\sigma_m,\eta)$. This is so since

$$\begin{aligned}\operatorname{Var}\big[\mathbb{E}_\eta U(\mathcal{D}_m^k)|\sigma_m\big] &= \mathbb{E}\big(\big\{\mathbb{E}_\eta\big[U(\mathcal{D}_m^k) - \mathbb{E}(U(\mathcal{D}_m^k)|\sigma_m,\eta)\big]\big\}^2|\sigma_m\big)\\ &\le \mathbb{E}\big(\mathbb{E}_\eta\big\{\big[U(\mathcal{D}_m^k) - \mathbb{E}(U(\mathcal{D}_m^k)|\sigma_m,\eta)\big]\big\}^2|\sigma_m\big)\\ &= \mathbb{E}_\eta\mathbb{E}\big(\big\{\big[U(\mathcal{D}_m^k) - \mathbb{E}(U(\mathcal{D}_m^k)|\sigma_m,\eta)\big]\big\}^2|\sigma_m,\eta\big)\\ &= \mathbb{E}_\eta\big\{\operatorname{Var}\big[U(\mathcal{D}_m^k)|\sigma_m,\eta\big]\big\},\end{aligned}$$

where we used the fact that η is independent of all other randomness. Now from Lemma 5.1 of Neykov et al. [2021] we can conclude

$$\sum_{m\in[d]} \mathrm{Var}(T^m|\sigma_m) \le (\log_2 d) \sum_{m\in[d]} \sigma_m^2 \mathbb{1}(\sigma_m \ge 4) C \sum_k \frac{1}{2^{4k}} \mathbb{E}_\eta \left(\frac{\mathbb{E}[U(\mathcal{D}_m^k)|\sigma_m, \eta]}{\sigma_m} + \frac{1}{\sigma_m^2} \right)$$
$$\le (\log_2 d) \sum_{m\in[d]} C \left(\sum_k 1/2^{2k} \mathbb{E}[U(\mathcal{D}_m^k)\sigma_m|\sigma_m] + \mathbb{1}(\sigma_m \ge 4) \right).$$

Taking expectation of the expression above we end up with

$$\mathbb{E}[\mathrm{Var}[T|\sigma]] \le C(\log_2 d) \left(\mathbb{E}[T] + \mathbb{E} \sum_{m\in[d]} \mathbb{1}(\sigma_m \ge 4) \right) \le C(\log_2 d)(\mathbb{E}[T] + d).$$

For the second term we have

$$\mathbb{E}[T|\sigma] = \sum_{m\in[d]} \sigma_m \mathbb{1}(\sigma_m \ge 4) \sum_k 1/2^{2k} \mathbb{E}[\mathbb{E}_\eta[U(\mathcal{D}_m^k)]|\sigma_m]$$
$$= \sum_{m\in[d]} \sigma_m \mathbb{1}(\sigma_m \ge 4) \sum_k 1/2^{2k} \sum_{i,j} \mathbb{E}_\eta (q_{ij}^k(m) - q_{i\cdot}^k(m) q_{\cdot j}^k(m))^2$$

Since the σ_m are independent we have

$$\mathrm{Var}[\mathbb{E}[T|\sigma]] = \sum_{m\in[d]} \mathrm{Var}[\sigma_m \mathbb{1}(\sigma_m \ge 4)] \left(\sum_k 1/2^{2k} \mathbb{E}_\eta \sum_{i,j} (q_{ij}^k(m) - q_{i\cdot}^k(m) q_{\cdot j}^k(m))^2 \right)^2$$

By Claim 2.1 of Canonne et al. [2018], we have that $\mathrm{Var}[\sigma_m \mathbb{1}(\sigma_m \ge 4)] \le C' \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \ge 4)]$, and $\sum_{i,j} (q_{ij}^k(m) - q_{i\cdot}^k(m) q_{\cdot j}^k(m))^2 \le (\sum_{i,j} |q_{ij}^k(m) - q_{i\cdot}^k(m) q_{\cdot j}^k(m)|)^2 \le 4$ thus

$$\mathrm{Var}[\mathbb{E}[T|\sigma]] \le 4C' \sum_{m\in[d]} \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \ge 4)] \sum_k 1/2^{2k} \mathbb{E}_\eta \sum_{i,j} (q_{ij}^k(m) - q_{i\cdot}^k(m) q_{\cdot j}^k(m))^2 = 4C' \mathbb{E}[T].$$

Hence $\mathrm{Var}\, T \le (\log_2 d) C (\mathbb{E}[T] + d)$.

2.3 Putting Things Together

Recall that the threshold $\tau = \zeta \sqrt{d} (\log_2 d)^2$ while $d \asymp n^{2/5}$. First we handle the null hypothesis. By Chebyshev's inequality we have

$$\mathbb{P}(|T - \mathbb{E}T| \ge \frac{\tau}{2}) \le \frac{4\,\mathrm{Var}(T)}{\tau^2} = \frac{C \log_2 d (\mathbb{E}[T] + d)}{\tau^2} \le \frac{C \log_2 d (\frac{C(\log_2 d)^2 n}{d^2} + d)}{\zeta d \log_2^4 d} \le \frac{1}{10},$$

when ζ is large enough. In this scenario we have that $T \le \frac{\tau}{2} + \mathbb{E}T$ which is of the order $\frac{Cn}{d^2} + \sqrt{\zeta d \log_2^4 d}/2$. Under the alternative, as we argued in (2.8): $\mathbb{E}[T] \gtrsim \min\left\{ \frac{n\Upsilon^2}{\lceil \log_2 d \rceil}, \frac{(n\Upsilon)^4}{d^3 \lceil \log_2 d \rceil^2} \right\}$. When $\Upsilon \ge \frac{(\log_2 d)^{3/2}}{d}$, simple algebra (using $d \asymp n^{2/5}$) shows that $\min\left\{ \frac{n\Upsilon^2}{\lceil \log_2 d \rceil}, \frac{(n\Upsilon)^4}{d^3 \lceil \log_2 d \rceil^2} \right\} = \frac{n\Upsilon^2}{\lceil \log_2 d \rceil}$, so that

$$\mathbb{P}(|T - \mathbb{E}T| \ge \mathbb{E}T/2) \le \frac{4\,\mathrm{Var}\, T}{(\mathbb{E}T)^2} \le 4C \left(\frac{d \log_2 d}{(\mathbb{E}T)^2} + \frac{\log_2 d}{\mathbb{E}T} \right) \le \frac{1}{10},$$

since $\mathbb{E}T \ge \zeta \sqrt{d} \log_2^2 d$ in order when $d \asymp n^{2/5}$.

3 Lower Bound

In this section we consider a lower bound which nearly matches the upper bound from Theorem 2.3. The main result of this section is as follows.

Theorem 3.1 (Critical Radius Lower Bound)**.** Let $L \in \mathbb{R}^+$ be a fixed constant. Then for some absolute constant $c_0 > 0$ the critical radius defined in (1.3) is bounded as

$$\varepsilon_n(\mathcal{P}_0^W(L), \mathcal{P}_1^W(L, \varepsilon)) \geq \frac{c_0}{n^{1/5}}.$$

One can see that the minimax rate given by Theorem 3.1 nearly matches (up to logarithmic factors) the rate from the upper bound of Theorem 2.3. Here once again we would like to stress the fact that the minimax optimal rate obtained here under W_1 smoothness and W_2 separation is slower compared to the rates obtained by Neykov et al. [2021]. Hence even though the separation is stronger (see Proposition 1.5) the added flexibility from only imposing Wasserstein smoothness drives the slower rate. The remaining of this section is dedicated to the proof of the above theorem. The techniques we use are similar to those used by Neykov et al. [2021], but the worst case is quite different. Unlike in Neykov et al. [2021] where the authors considered cases where $p_{X,Y|Z}$ are discrete or continuous distributions and obtained different rates, here there is no need for that. The worst case is achieved by a distribution which is discrete $X, Y|Z = z$ for all $z \in [0, 1]$, and in fact is concentrated only on 4 points. Before we detail this construction we will require the following lemma, which is useful when we establish the W_2 separation in the alternative.

Lemma 3.2. For any distribution $p = p_{X,Y,Z}$ for $X, Y, Z \in [0, 1]$, $\psi(p) \leq \widetilde{\psi}(p) \leq C\psi(p)$ for some absolute constant C where

$$\widetilde{\psi}(p) = \int W_2(p_{XY|Z=z}, p_{X|Z=z} \times p_{Y|Z=z}) dp_Z(z),$$

and

$$\psi(p) = \inf_{q \in \mathcal{P}_0} \int W_2(p_{XY|Z=z}, q_{X|Z=z} \times q_{Y|Z=z}) dp_Z(z). \tag{3.1}$$

Proof. Let $p^* = p^*_{X|Z} \times p^*_{Y|Z}$ be the minimizer of $\psi(p)$ (if a minimizer does not exist we may take a sequence of distributions that converges to it). Then

$$\begin{aligned}
\widetilde{\psi}(p) &= \int W_2(p_{XY|Z=z}, p_{X|Z=z} \times p_{Y|Z=z}) dp_Z(z) \\
&\leq \int W_2(p_{XY|Z=z}, p^*_{X|Z=z} \times p^*_{Y|Z=z}) dp_Z(z) \\
&\quad + \int W_2(p_{X|Z=z} \times p_{Y|Z=z}, p^*_{X|Z=z} \times p^*_{Y|Z=z}) dp_Z(z) \\
&\overset{\text{by Lemma 3 Mariucci and Reiß [2018]}}{\leq} \psi(p) + \int \sqrt{[W_2^2(p_{X|Z=z}, p^*_{X|Z=z}) + W_2^2(p_{Y|Z=z}, p^*_{Y|Z=z})]} dp_Z(z) \\
&\leq \psi(p) + \int [W_2(p_{X|Z=z}, p^*_{X|Z=z}) + W_2(p_{Y|Z=z}, p^*_{Y|Z=z})] dp_Z(z) \\
&\leq \psi(p) + \int W_2(p_{XY|Z=z}, p^*_{X|Z=z} \times p^*_{Y|Z=z}) dp_Z(z) + \int W_2(p_{XY|Z=z}, p^*_{X|Z=z} \times p^*_{Y|Z=z}) dp_Z(z) \\
&= C\psi(p).
\end{aligned}$$

In the above we used that

$$\int W_2(p_{X|Z=z}, p^*_{X|Z=z}) dp_Z(z) \leq \int W_2(p_{XY|Z=z}, p^*_{X|Z=z} \times p^*_{Y|Z=z}) dp_Z(z),$$

which follows by similar arguments as in Lemma 2.4. □

Let Z be $U([0,1])$. For the null distributions for each z we specify $q_{X,Y|Z=z} = q_{X|Z=z}q_{Y|Z=z}$ as four point masses at $(0,0),(0,1),(1,0)$ and $(1,1)$ with equal probability $\frac{1}{4}$. Under the alternative we specify $p_{X,Y|Z=z}$ as four point masses at $(0,0),(0,1),(1,0)$ and $(1,1)$, where $p_{X,Y|Z=z}(0,0) = p_{X,Y|Z=z}(1,1) = \frac{1}{4} + \delta\xi(z)$ and $p_{X,Y|Z=z}(0,1) = p_{X,Y|Z=z}(1,0) = \frac{1}{4} - \delta\xi(z)$, where $\xi(z)$ is specified as:

$$\xi_\nu(z) = \rho \sum_{j\in[d]} \nu_j h_{j,d}(z),$$

where $\rho > 0$ is a constant, $\delta \in \{-1,1\}$ is a Rademacher random variable, $d \in \mathbb{N}$, $\nu_i \in \{-1,+1\}$, and $h_{j,d}(z) = \sqrt{d}h(dz-j+1)$ for $z \in [(j-1)/d, j/d]$, and h is an infinitely differentiable function supported on $[0,1]$

such that $\int h(z)dz = 0$ and $\int h^2(z)dz = 1$. When perturbing, in order to ensure that we create valid probability distributions, we need to satisfy the conditions that

$$\frac{1}{4} - \rho\sqrt{d}\|h\|_\infty \geq 0, \quad \frac{1}{4} + \rho\sqrt{d}\|h\|_\infty \leq 1. \tag{3.2}$$

Clearly, the second inequality in (3.2) is implied by the first one and is hence redundant. We will ensure the first inequality by the choice of ρ and d to follow.

We need to show that the two distributions q and p are Wasserstein smooth. This is obvious for q which does not change with z. To see this for $p_{X,Y|Z=z}$ and $p_{X,Y|Z=z'}$ we will use the dual characterization of the Wasserstein distance. We have

$$\begin{aligned} W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) &= \sup_{f\in\text{Lip}(1)} |f(0,0) + f(1,1) - f(0,1) - f(1,0)||\xi(z) - \xi(z')| \\ &\leq 2|\xi(z) - \xi(z')| \\ &\leq 2d^{3/2}\rho\|h'\|_\infty|z - z'|, \end{aligned}$$

since the derivative of $\xi(z)$ is bounded by $d^{3/2}\rho\|h'\|_\infty$. Next we will handle $\psi(p)$ as defined in (3.1). We will start by checking that $W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z})$ is at least $C|\xi(z)|$ for some constant C. Once again we use the dual characterization of the Wasserstein distance to obtain

$$W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) = \sup_{f\in\text{Lip}(1)} |f(0,0) + f(1,1) - f(0,1) - f(1,0)||\xi(z)|$$

Take the function $f(x,y) = \frac{1}{\sqrt{2}}|x-y|$ where $x, y \in \mathbb{R}$. This is a 1-Lipschitz function in $\|\cdot\|_2$ (since $||x-y| - |x'-y'|| \leq |x-x'+y'-y| \leq |x-x'| + |y-y'| \leq \sqrt{2((x-x')^2 + (y-y')^2)}$). It follows that

$$W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) \geq \sqrt{2}|\xi(z)|.$$

Next we show that $W_2^2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) \asymp W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z})$. This follows since when $x, y \in \{(0,0),(1,1),(0,1),(1,0)\}$ we have

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

Hence if γ_2 is an optimal coupling for $W_2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z})$ we have

$$\begin{aligned} W_2^2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) &= \int \|x-y\|_2^2 d\gamma_2(x,y) \\ &\geq \int \|x-y\|_2 d\gamma_2(x,y) \geq W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}). \end{aligned}$$

On the other hand if γ_1 is an optimal coupling for $W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z})$ we have

$$W_2^2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) \leq \int \|x-y\|_2^2 d\gamma_1(x,y)$$
$$\leq \sqrt{2}\int \|x-y\|_2 d\gamma_1(x,y) = \sqrt{2}W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}). \tag{3.3}$$

As we argued earlier, $\sqrt{W_1(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z})} \gtrsim \sqrt{|\xi(z)|} = \sqrt{\rho|\sum_{j\in[d]}\nu_j h_{j,d}(z)|}$, where $\rho > 0$ is a constant, $d \in \mathbb{N}$, $\nu_i \in \{-1,+1\}$,and $h_{j,d}(z) = \sqrt{d}h(dz - j + 1)$ for $z \in [(j-1)/d, j/d]$. Now, since the functions $h_{j,d}$ have disjoint supports, it follows that $\sqrt{\rho|\sum_{j\in[d]}\nu_j h_{j,d}(z)|} = \sqrt{\rho}\sum_{j\in[d]}\sqrt{|h_{j,d}(z)|}$. Hence $\mathbb{E}\sqrt{|\xi(z)|} = \mathbb{E}\sqrt{\rho}\sum_{j\in[d]}\sqrt{|h_{j,d}(Z)|} = d\sqrt{\rho}\sqrt[4]{d}\frac{1}{d} = c\sqrt{\rho}\sqrt[4]{d}$, where we used the fact that $\int\sqrt{|h_{j,d}(z)|}dz = c\sqrt[4]{d}\frac{1}{d}$, for some absolute constant c. We conclude that

$$\varepsilon = \psi(p) \gtrsim \rho^{1/2}d^{1/4}.$$

From here the argument can proceed precisely as in Theorem 4.1 of Neykov et al. [2021] where $\ell_1 = \ell_2 = 2$. We conclude that one can select $\rho \asymp \frac{1}{d^{3/2}}$ and $\frac{1}{d} \asymp \frac{1}{n^{2/5}}$ (for some sufficiently small constants so that (3.2) are satisfied), to yield a lower bound on $\psi(p) \gtrsim \frac{1}{\sqrt{d}} \gtrsim \frac{1}{n^{1/5}}$. This completes the proof.

4 Discussion

In this paper we considered the problem of minimax Wasserstein conditional independence testing. We proposed a novel test statistic which is nearly optimal in terms of the separation radius. Despite this, there are interesting open questions that remain to be explored. Our current theory allows only for 1-dimensional random variables X, Y, Z. It will be interesting (yet also very challenging) to extend our results to the multivariate setting. Furthermore, while in principle our test statistic is implementable in polynomial time, the computational cost which is bigger than linear is likely high. It would be interesting to design fast computational methods to compute our statistic, or propose a statistic which is different in nature altogether yet is minimax optimal and easily computable. Another challenging open question is whether one can come up with a calibration method for the test statistic such as the one proposed in Kim et al. [2022b] which is based on local permutations. The difficulty here is the fact that Wasserstein smoothness is not strong enough to apply a result such as Lemma 1 or Lemma 2 of Kim et al. [2022b], which renders it almost impossible to argue directly that a local permutation would control the type I error. Finally, while we have addressed minimax testing, it would be interesting to study the corresponding estimation problem, possibly under W_2 loss function with W_1 smoothness. We leave these important questions for future work.

5 Acknowledgements

This work was partially supported by funding from the NSF grants DMS2113684 and DMS-2310632, as well as an Amazon AI and a Google Research Scholar Award to SB. MN acknowledges support from the NSF grant DMS-2113684.

References

Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on hsic measures. *The Annals of Statistics*, 50(2):858–879, 2022.

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

Khanh Do Ba, Huy L. Nguyen, Huy N. Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, 48(2):428–442, 2011.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749, 2018.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, 47(4):1893–1927, 2019.

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

Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–57. IEEE, 2018.

Alexandra Carpentier and Nicolas Verzelen. Optimal sparsity testing in linear regression model. *Bernoulli*, 27(2):727–750, 2021.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Lucas De Lara, Alberto González-Sanz, and Jean-Michel Loubes. A consistent extension of discrete optimal transport maps for machine learning applications. *arXiv preprint arXiv:2102.08644*, 2021.

Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, pages 1–16, 2021.

Nabarun Deb, Bhaswar B Bhattacharya, and Bodhisattva Sen. Efficiency lower bounds for distribution-free hotelling-type two-sample tests based on optimal transport. *arXiv preprint arXiv:2104.01986*, 2021.

Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

Chris Finlay, Augusto Gerolin, Adam M Oberman, and Aram-Alexandre Pooladian. Learning normalizing flows from entropy-kantorovich potentials. *arXiv preprint arXiv:2006.06033*, 2020.

Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *arXiv preprint arXiv:2104.09418*, 2021.

Promit Ghosal and Bodhisattva Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50(2):1012–1037, 2022.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR, 2019.

Marc Hallin, Eustasio Del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd international workshop on statistical and computational theories of vision*, volume 2, page 5. Nice, France, 2003.

J.I. Ingster and I.A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Lecture Notes in Statistics. Springer, 2003.

Yuri Izmailovich Ingster. On the minimax nonparametric detection of signals in white gaussian noise. *Problemy Peredachi Informatsii*, 18(2):61–73, 1982.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251, 2022a.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022b.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Conditional Independence Testing for Discrete Distributions: Beyond χ^2- and G-tests. *arXiv preprint arXiv:2308.05373*, 2023.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Patrick T Komiske, Radha Mastandrea, Eric M Metodiev, Preksha Naik, and Jesse Thaler. Exploring the space of jets with cms open data. *Physical Review D*, 101(3):034009, 2020.

Oleg V Lepski and Vladimir G Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

Peihua Li, Qilong Wang, and Lei Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1696, 2013.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps, 2021.

Dimitris Margaritis. Distribution-free learning of bayesian network structure in continuous domains. In *AAAI*, volume 5, pages 825–830, 2005.

Ester Mariucci and Markus Reiß. Wasserstein and total variation distance between marginals of lévy processes. *Electronic Journal of Statistics*, 12(2):2482–2514, 2018.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232, 2021.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4):691–717, 2019.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in rd: an approach based on mixture modeling and optimal transport. *CoRR*, 2022.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Andrew Warren. Wasserstein conditional independence testing. *arXiv preprint arXiv:2107.14184*, 2021.

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

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, USA, 2011. AUAI Press.

A Deferred Proofs

Proof of Fact 1.2.

1. On bounded domains $W_2(p,q) \lesssim \mathrm{TV}(p,q)$ [See Lemma 3 and Theorem 6.15 Slawski and Sen, 2022, Villani, 2009, respectively] where $\lesssim$ denotes inequality up to absolute constant factors.
2. This result can be found in Equation (6.4) of Villani [2009].
3. The proof of this result can be found on page 77 of Villani [2009].
4. This follows from Lemma 3 of Mariucci and Reiß [2018].
5. This follows as in (3.3).

□

Proof of Lemma 2.1. Consider first $\widetilde{W}_2^2(p,q) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int \frac{\|x-y\|_2^2}{2} d\gamma(x,y)$. The grid Q^k with side Euclidean lengths (mostly) $\frac{1}{2^k}$ centered at the point η forms a dyadic partitioning (see Definition 1 [Weed and Bach, 2019]) for the scaled norm $\|x-y\|_2/\sqrt{2}$. The only condition that we need to check is whether $\mathrm{diam}(Q) \le \gamma^k$ for any $Q \in Q^k$. Since $Q \in Q^k$, then we have for $(x,y) \in Q$: $\|x-y\|_2^2 \le \frac{2}{2^{2k}}$ and hence $\frac{\|x-y\|_2^2}{2} \le \frac{1}{2^{2k}}$, so it is a dyadic partitioning with $\gamma = \frac{1}{2}$. By Proposition 1 of Weed and Bach [2019] we immediately conclude that

$$\widetilde{W}_2^2(p,q) \le \frac{1}{2^{2\lceil \log_2(d)\rceil}} + \sum_{k=1}^{\lceil \log_2(d)\rceil} \frac{1}{2^{2(k-1)}} \sum_{A_{ij}^k \in Q^k} |p(A_{ij}^k) - q(A_{ij}^k)|.$$

Multiplying back by 2 on both sides proves the desired result (with a multiplicative constant 4). □

Proof of Lemma 2.4. The first inequality is true by assumption. The second inequality simply plugs in the distribution $p_{X|Z=z}p_{Y|Z=z}$ in place of $q_{X,Y|Z=z}$. We now prove the last inequality.

We will first show that the distributions $p_{X,Y|Z\in C_m}$ and $p_{X,Y|Z=z'}$ for any $z' \in C_m$ are close in the W_2 distance. To see this, fix a z' and suppose that $\gamma_z(x,y,x',y')$ is an optimal coupling between $P_{X,Y|Z}(x,y|z')$ and $P_{X,Y|Z}(x,y|z)$ which minimizes the Wasserstein distance

$$W_2^2(p_{X,Y|Z=z'}, p_{X,Y|Z=z}) = \int \|(x,y)-(x',y')\|_2^2 d\gamma_z(x,y,x',y'),$$

and satisfies $\gamma_z(x,y,\infty,\infty) = P_{X,Y|Z}(x,y|z')$ and $\gamma_z(\infty,\infty,x',y') = P_{X,Y|Z}(x',y'|z)$. Such an optimal coupling exists due to Theorem 4.1 of Villani [2009]. Take the mixture of such distributions over z, i.e., consider the coupling $\int_{C_m} \gamma_z(x,y,x',y') d\widetilde{P}(z)$, where $d\widetilde{P}(z) = dP(z)/\mathbb{P}(Z \in C_m)$. Note that this is a coupling between the distributions $P_{X,Y|Z}(x,y|z')$ and $P_{X,Y|Z}(x,y|z \in C_m)$, since $\int_{C_m} \gamma_z(x,y,\infty,\infty) d\widetilde{P}(z) = P_{X,Y|Z}(x,y|z')$ and

$$\int_{C_m} \gamma_z(\infty,\infty,x',y') d\widetilde{P}(z) = \int_{C_m} P_{X,Y|Z}(x',y'|z) d\widetilde{P}(z) = P_{X,Y|Z\in C_m}(x',y'|z \in C_m).$$

Now consider

$$\begin{aligned} W_2^2(p_{X,Y|Z\in C_m}, p_{X,Y|Z=z'}) &\le \int \|(x,y)-(x',y')\|_2^2 d \int_{C_m} \gamma_z(x,y,x',y') d\widetilde{P}(z) \\ &= \int_{C_m} \int \|(x,y)-(x',y')\|_2^2 d\gamma_z(x,y,x',y') d\widetilde{P}(z) \\ &= \int_{C_m} W_2^2(p_{X,Y|Z=z'}, p_{X,Y|Z=z}) d\widetilde{P}(z) \\ &\lesssim L\,\mathrm{diam}(C_m), \end{aligned} \tag{A.1}$$

where the last inequality follows from the fact that $W_1 \gtrsim W_2^2$, since we are on a bounded domain and using Assumptions 1.3 and 1.4.

We will now show that $W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \geq W_1(p_{X|Z=z}, p_{X|Z=z'})$. Let $\gamma(x, y, x', y')$ denote an optimal coupling between $p_{X,Y|Z=z}$ and $p_{X,Y|Z=z'}$. Note that $\gamma(x, \infty, x', \infty) = \int_{y,y'} d\gamma(x, y, x', y')$ is a coupling between $p_{X|Z=z}$ and $p_{X|Z=z'}$. Thus we have

$$\begin{aligned} W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) &= \int \|(x, y) - (x', y')\|_2 d\gamma(x, y, x', y') \\ &\geq \int \|(x, 0) - (x', 0)\|_2 d\gamma(x, \infty, x', \infty) \\ &\geq W_1(p_{X|Z=z}, p_{X|Z=z'}). \end{aligned} \tag{A.2}$$

Similarly, one can argue that $W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \geq W_1(p_{Y|Z=z}, p_{Y|Z=z'})$. Hence by the same reasoning as in (A.1) we may show that under the condition $W_1(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \leq L|z - z'|$ we have

$$\max\big(W_2^2(p_{X|Z=z}, p_{X|Z\in C_m}), W_2^2(p_{Y|Z=z}, p_{Y|Z\in C_m})\big) \lesssim L \operatorname{diam}(C_m).$$

Next we will show that for any $z \in C_m$ we have

$$|W_2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) - W_2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m}p_{Y|Z\in C_m})| \lesssim \big(L \operatorname{diam}(C_m)\big)^{1/2}.$$

First we observe that:

$$W_2(p_{X|Z=z}p_{Y|Z=z}, p_{X|Z\in C_m}p_{Y|Z\in C_m}) \leq \sqrt{W_2^2(p_{X|Z=z}, p_{X|Z\in C_m}) + W_2^2(p_{Y|Z=z}, p_{Y|Z\in C_m})}$$

where we used Lemma 3 of Mariucci and Reiß [2018], which shows that the Wasserstein distance squared with a distance function equal to $\|\cdot\|_2$ norm, is sub-additive on product distributions. Next, by the triangle inequality we have

$$\begin{aligned} &|W_2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) - W_2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m}p_{Y|Z\in C_m})| \\ &\quad\leq W_2(p_{X,Y|Z=z}, p_{X,Y|Z\in C_m}) + W_2(p_{X|Z=z}p_{Y|Z=z}, p_{X|Z\in C_m}p_{Y|Z\in C_m}) \\ &\quad\leq W_2(p_{X,Y|Z=z}, p_{X,Y|Z\in C_m}) + \sqrt{W_2^2(p_{X|Z=z}, p_{X|Z\in C_m}) + W_2^2(p_{Y|Z=z}, p_{Y|Z\in C_m})} \\ &\quad\lesssim \big(L \operatorname{diam}(C_m)\big)^{1/2}. \end{aligned}$$

Integrating the above inequalities over z we obtain

$$\begin{aligned} \int_{C_m} W_2(p_{X,Y|Z=z}, p_{X|Z=z}p_{Y|Z=z}) dP(z) \leq [&W_2(p_{X,Y|Z\in C_m}, p_{X|Z\in C_m}p_{Y|Z\in C_m}) \\ &+ \kappa(L \operatorname{diam}(C_m))^{1/2}]p_m, \end{aligned}$$

where $\kappa > 0$ is an absolute constant. Summing up over m gives the inequality that we wanted to show. □