A Higher-Order Kolmogorov-Smirnov Test

Veeranjaneyulu Sadhanala[1] **Yu-Xiang Wang**[2] **Aaditya Ramdas**[1] **Ryan J. Tibshirani**[1]
[1]Carnegie Mellon University [2]University of California at Santa Barbara

Abstract

We present an extension of the Kolmogorov-Smirnov (KS) two-sample test, which can be more sensitive to differences in the tails. Our test statistic is an integral probability metric (IPM) defined over a higher-order total variation ball, recovering the original KS test as its simplest case. We give an exact representer result for our IPM, which generalizes the fact that the original KS test statistic can be expressed in equivalent variational and CDF forms. For small enough orders ($k \leq 5$), we develop a linear-time algorithm for computing our higher-order KS test statistic; for all others ($k \geq 6$), we give a nearly linear-time approximation. We derive the asymptotic null distribution for our test, and show that our nearly linear-time approximation shares the same asymptotic null. Lastly, we complement our theory with numerical studies.

1 INTRODUCTION

The Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1948) is a classical and celebrated tool for nonparametric hypothesis testing. Let $x_1, \ldots, x_m \sim P$ and $y_1, \ldots, y_n \sim Q$ be independent samples. Let $X_{(m)}$ and $Y_{(n)}$ denote the two sets of samples, and also let $Z_{(N)} = X_{(m)} \cup Y_{(n)} = \{z_1, \ldots, z_N\}$, where $N = m+n$. The two-sample KS test statistic is defined as

$$\max_{z \in Z_{(m+n)}} \left| \frac{1}{m} \sum_{i=1}^m 1\{x_i \leq z\} - \frac{1}{n} \sum_{i=1}^n 1\{y_i \leq z\} \right|. \quad (1)$$

In words, this measures the maximum absolute difference between the empirical cumulative distribution functions (CDFs) of $X_{(m)}$ and $Y_{(n)}$, across all points in the joint sample $Z_{(m+n)}$. Naturally, the two-sample KS test rejects the null hypothesis of $P = Q$ for large values of the statistic. The statistic (1) can also be written in the following variational form:

$$\sup_{f \,:\, \mathrm{TV}(f) \leq 1} |\mathbb{P}_m f - \mathbb{Q}_n f|, \quad (2)$$

where $\mathrm{TV}(\cdot)$ denotes total variation, and we define the empirical expectation operators $\mathbb{P}_m, \mathbb{Q}_n$ via

$$\mathbb{P}_m f = \frac{1}{m} \sum_{i=1}^m f(x_i) \;\text{ and }\; \mathbb{Q}_n f = \frac{1}{n} \sum_{i=1}^n f(y_i).$$

Later, we will give a general representation result that implies the equivalence of (1) and (2) as a special case.

The KS test is a fast, general-purpose two-sample nonparametric test. But being a general-purpose test also means that it is systematically less sensitive to some types of differences, such as tail differences (Bryson, 1974). Intuitively, this is because the empirical CDFs of $X_{(m)}$ and $Y_{(n)}$ must both tend to 0 as $z \to -\infty$ and to 1 as $z \to \infty$, so the gap in the tails will not be large.

The insensitivity of the KS test to tail differences is well-known. Several authors have proposed modifications to the KS test to improve its tail sensitivity, based on variance-reweighting (Anderson and Darling, 1952), or Renyi-type statistics (Mason and Schuenemeyer, 1983; Calitz, 1987), to name a few ideas. In a different vein, Wang et al. (2014) recently proposed a higher-order extension of the KS two-sample test, which replaces the total variation constraint on f in (2) with a total variation constraint on a derivative of f. These authors show empirically that, in some cases, this modification can lead to better tail sensitivity. In the current work, we refine the proposal of Wang et al. (2014), and give theoretical backing for this new test.

A Higher-Order KS Test. Our test statistic has the form of an integral probability metric (IPM). For a function class $\mathcal{F}$, the IPM between distributions P and Q, with respect to $\mathcal{F}$, is defined as (Muller, 1997)

$$\rho(P, Q; \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{P} f - \mathbb{Q} f| \quad (3)$$

where we define the expectation operators $\mathbb{P}, \mathbb{Q}$ by

$$\mathbb{P} f = \mathbb{E}_{X \sim P}[f(X)] \;\text{ and }\; \mathbb{Q} f = \mathbb{E}_{Y \sim Q}[f(Y)].$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

For a given function class $\mathcal{F}$, the IPM $\rho(\cdot, \cdot; \mathcal{F})$ is a pseudometric on the space of distributions. Note that the KS test in (2) is precisely $\rho(P_m, Q_n; \mathcal{F}_0)$, where P_m, Q_n are the empirical distributions of $X_{(m)}, Y_{(n)}$, respectively, and $\mathcal{F}_0 = \{f : \mathrm{TV}(f) \le 1\}$.

Consider an IPM given by replacing $\mathcal{F}_0$ with $\mathcal{F}_k = \{f : \mathrm{TV}(f^{(k)}) \le 1\}$, for an integer $k \ge 1$ (where we write $f^{(k)}$ for the kth weak derivative of f). Some motivation is as follows. In the case $k = 0$, we know that the *witness functions* in the KS test (2), i.e., the functions in $\mathcal{F}_0$ that achieve the supremum, are piecewise constant step functions (cf. the equivalent representation (1)). These functions can only have so much action in the tails. By moving to $\mathcal{F}_k$, which is essentially comprised of the kth order antiderivative of functions in $\mathcal{F}_0$, we should expect that the witness functions over $\mathcal{F}_k$ are kth order antiderivatives of piecewise constant functions, i.e., kth degree piecewise polynomial functions, which can have much more sensitivity in the tails.

But simply replacing $\mathcal{F}_0$ by $\mathcal{F}_k$ and proposing to compute $\rho(P_m, Q_n; \mathcal{F}_k)$ leads to an ill-defined test. This is due to the fact that $\mathcal{F}_k$ contains all polynomials of degree k. Hence, if the ith moments of P_m, Q_n differ, for any $i \in [k]$ (where we abbreviate $[a] = \{1, \dots, a\}$ for an integer $a \ge 1$), then $\rho(P_m, Q_n; \mathcal{F}_k) = \infty$.

As such, we must modify $\mathcal{F}_k$ to control the growth of its elements. While there are different ways to do this, not all result in computable IPMs. The approach we take yields an exact representer theorem (generalizing the equivalence between (1) and (2)). Define

$$\mathcal{F}_k = \big\{ f : \mathrm{TV}(f^{(k)}) \le 1, \\ f^{(j)}(0) = 0,\ j \in \{0\} \cup [k-1], \\ f^{(k)}(0+) = 0 \text{ or } f^{(k)}(0-) = 0 \big\}. \quad (4)$$

Here $f^{(k)}(0+)$ and $f^{(k)}(0-)$ denote one-sided limits at 0 from above and below, respectively. Informally, the functions in $\mathcal{F}_k$ are pinned down at 0, with all lower-order derivatives (and the limiting kth derivative from the right or left) equal to 0, which limits their growth. Now we define the *kth-order KS test statistic* as

$$\rho(P_m, Q_n; \mathcal{F}_k) = \sup_{f \in \mathcal{F}_k} |\mathbb{P}_m f - \mathbb{Q}_n f|. \quad (5)$$

An important remark is that for $k = 0$, this recovers the original KS test statistic (2), because $\mathcal{F}_0$ contains all step functions of the form $g_t(x) = 1\{x \le t\}$, $t \ge 0$.

Another important remark is that for any $k \ge 0$, the function class $\mathcal{F}_k$ in (4) is "rich enough" to make the IPM in (5) a metric. We state this formally next; its proof, as with all other proofs, is in the appendix.

Proposition 1. *For any $k \ge 0$, and any P, Q with k moments, $\rho(P, Q; \mathcal{F}_k) = 0$ if and only if $P = Q$.*

Motivating Example. Figure 1 shows the results of a simple simulation comparing the proposed higher-order tests (5), of orders $k = 1$ through 5, against the usual KS test (corresponding to $k = 0$). For the simulation setup, we used $P = N(0, 1)$ and $Q = N(0, 1.44)$. For 500 "alternative" repetitions, we drew $m = 250$ samples from P, drew $n = 250$ samples from Q, and computed test statistics; for another 500 "null" repetitions, we permuted the $m + n = 500$ samples from the corresponding alternative repetition, and again computed test statistics. For each test, we varied the rejection threshold for each test, we calculated its true positive rate using the alternative repetitions, and calculated its false positive rate using the null repetitions. The oracle ROC curve corresponds to the likelihood ratio test (which knows the exact distributions P, Q). Interestingly, we can see that power of the higher-order KS test improves as we increase the order from $k = 0$ up to $k = 2$, then stops improving by $k = 3, 4, 5$.

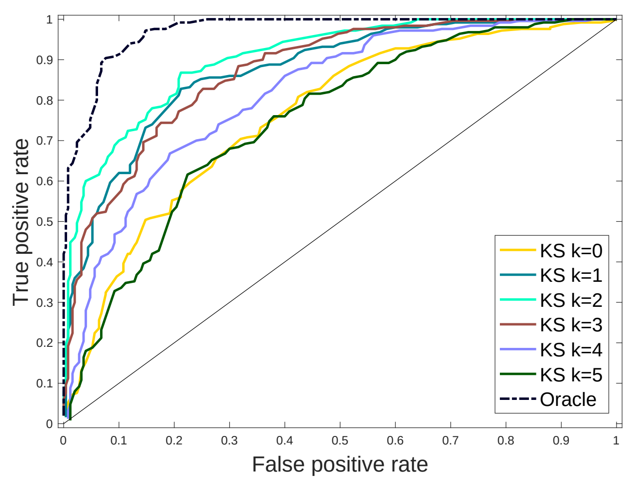


Figure 1: *ROC curves from an experiment comparing the proposed higher-order KS tests in* (5) *(for various k) to the usual KS test, when $P = N(0, 1)$ and $Q = N(0, 1.44)$.*

Figure 2 displays the witness function (which achieves the supremum in (5)) for a large-sample version of the higher-order KS test, across orders $k = 0$ through 5. We used the same distributions as in Figure 1, but now $n = m = 10^4$. We will prove in Section 2 that, for the kth order test, the witness function is always a kth degree piecewise polynomial (in fact, a rather simple one, of the form $g_t(x) = (x - t)_+^k$ or $g_t(x) = (t - x)_+^k$ for a knot t). Recall the underlying distributions P, Q here have different variances, and we can see from their witness functions that all higher-order KS tests choose to put weight on tail differences. Of course, the power of any test of is determined by the size of the statistic under the alternative, relative to typical fluctuations under the null. As we place more weight on tails, in this particular setting, we see diminishing returns at $k = 3$, meaning the null fluctuations must be too great.

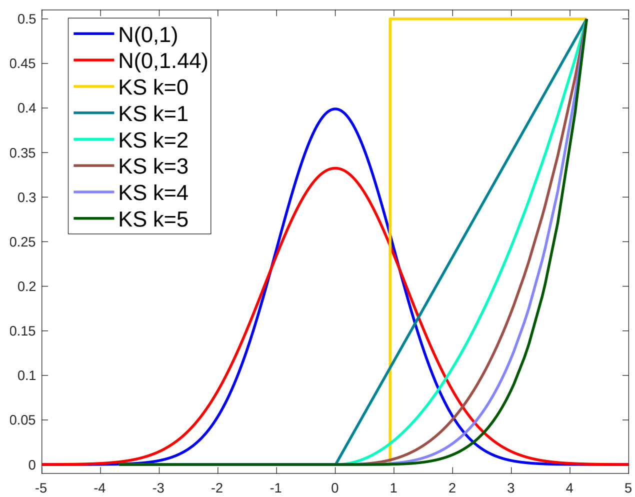


Figure 2: *Witness functions (normalized for plotting purposes) for the higher-order KS tests, when* $P = N(0,1)$ *and* $Q = N(0,1.44)$*. They are always of piecewise polynomial form; and here they all place weight on tail differences.*

Summary of Contributions. Our contributions in this work are as follows.

- We develop an exact representer theorem for the higher-order KS test statistic (5). This enables us to compute the test statistic in linear-time, for all $k \le 5$. For $k \ge 6$, we develop a nearly linear-time approximation to the test statistic.
- We derive the asymptotic null distribution of the our higher-order KS test statistic, based on empirical process theory. For $k \ge 6$, our approximation to the test statistic has the same asymptotic null.
- We provide concentration tail bounds for the test statistic. Combined with the metric property from Proposition 1, this shows that our higher-order KS test is asymptotically powerful against any pair of fixed, distinct distributions P, Q.
- We perform extensive numerical studies to compare the newly proposed tests with several others.

Other Related Work. Recently, IPMs have been gaining in popularity due in large part to energy distance tests (Szekely and Rizzo, 2004; Baringhaus and Franz, 2004) and kernel maximum mean discrepancy (MMD) tests (Gretton et al., 2012), and in fact, there is an equivalence between the two classes (Sejdinovic et al., 2013). An IPM with a judicious choice of $\mathcal{F}$ gives rise to a number of common distances between distributions, such as Wasserstein distance or total variation (TV) distance. While IPMs look at differences $dP - dQ$, tests based on ϕ-divergences (such as Kullback-Leibler, or Hellinger) look at ratios dP/dQ, but can be hard to efficiently estimate in practice (Sriperumbudur et al., 2009). The TV distance is the only IPM that is also a ϕ-divergence, but it is impossible to estimate.

There is also a rich class of nonparametric tests based on graphs. Using minimum spanning trees, Friedman and Rafsky (1979) generalized both the Wald-Wolfowitz runs test and the KS test. Other tests are based on k-nearest neighbors graphs (Schilling, 1986; Henze, 1988) or matchings (Rosenbaum, 2005). The Mann-Whitney-Wilcoxon test has a multivariate generalization using the concept of data depth (Liu and Singh, 1993). Bhattacharya (2016) established that many computationally efficient graph-based tests have suboptimal statistical power, but some inefficient ones have optimal scalings.

Different computational-statistical tradeoffs were also discovered for IPMs (Ramdas et al., 2015b). Further, as noted by Janssen (2000) (in the context of one-sample testing), every nonparametric test is essentially powerless in an infinity of directions, and has nontrivial power only against a finite subspace of alternatives. In particular, this implies that no single nonparametric test can uniformly dominate all others; improved power in some directions generally implies weaker power in others. This problem only gets worse in high-dimensional settings (Ramdas et al., 2015a; Arias-Castro et al., 2018). Therefore, the question of which test to use for a given problem must be guided by a combination of simulations, computational considerations, a theoretical understanding of the pros/cons of each test, and a practical understanding of the data at hand.

Outline. In Section 2, we give computational details for the higher-order KS test statistic (5). We derive its asymptotic null in Section 3, and give concentration bounds (for the statistic around the population-level IPM) in Section 4. We give numerical experiments in Section 5, and conclude in Section 6 with a discussion.

2 COMPUTATION

Write $T = \rho(P_m, Q_n; \mathcal{F}_k)$ for the test statistic in (5). In this section, we derive a representer theorem for T, develop a linear-time algorithm for $k \le 5$, and a nearly linear-time approximation for $k \ge 6$.

2.1 Representer Theorem

The higher-order KS test statistic in (5) is defined by an infinite-dimensional maximization over $\mathcal{F}_k$ in (4). Fortunately, we can restrict our attention to a simpler function class, as we show next.

Theorem 1. *Fix* $k \ge 0$*. Let* $g_t^+(x) = (x-t)_+^k/k!$ *and* $g_t^-(x) = (t-x)_+^k/k!$ *for* $t \in \mathbb{R}$*, where we write* $(a)_+ =$

$\max\{a, 0\}$. *For the statistic* T *defined by* (5),

$$T = \max\Big\{ \sup_{t \geq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+|, \sup_{t \leq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^-| \Big\}. \quad (6)$$

The proof of this theorem uses a key result from Mammen (1991), where it is shown that we can construct a spline interpolant to a given function at given points, such that its higher-order total variation is no larger than that of the original function.

Remark 1. When $k = 0$, note that for $t \geq 0$,

$$\begin{aligned} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+| &= \left| \frac{1}{m} \sum_{i=1}^m 1\{x_i > t\} - \frac{1}{n} \sum_{i=1}^n 1\{y_i > t\} \right| \\ &= \left| \frac{1}{m} \sum_{i=1}^m 1\{x_i \leq t\} - \frac{1}{n} \sum_{i=1}^n 1\{y_i \leq t\} \right| \end{aligned}$$

and similarly for $t \leq 0$, $|(\mathbb{P}_m - \mathbb{Q}_n) g_t^-|$ reduces to the same expression in the second line above. As we vary t from $-\infty$ to ∞, this only changes at values $t \in Z_{(N)}$, which shows (6) and (1) are the same, i.e., Theorem 1 recovers the equivalence between (2) and (1).

Remark 2. For general $k \geq 0$, we can interpret (6) as a comparison between truncated kth order moments, between the empirical distributions P_m and Q_n. The test statistic T the maximum over all possible truncation locations t. The critical aspect here is *truncation*, which makes the higher-order KS test statistic a metric (recall Proposition 1). A comparison of moments, alone, would not be enough to ensure such a property.

Theorem 1 itself does not immediately lead to an algorithm for computing T, as the range of t considered in the suprema is infinite. However, through a bit more work, detailed in the next two subsections, we can obtain an exact linear-time algorithm for all $k \leq 5$, and a linear-time approximation for $k \geq 6$.

2.2 Linear-Time Algorithm for $k \leq 5$

The key fact that we will exploit is that the criterion in (6), as a function of t, is a piecewise polynomial of order k with knots in $Z_{(N)}$. Assume without a loss of generality that $z_1 < \cdots < z_N$. Also assume without a loss of generality that $z_1 \geq 0$ (this simplifies notation, and the general case follows by the repeating the same arguments separately for the points in $Z_{(N)}$ on either side of 0). Define $c_i = \mathbb{1}\{z_i \in X_{(m)}\}/m - \mathbb{1}\{z_i \in Y_{(n)}\}/n$, $i \in [N]$, and

$$\phi_i(t) = \frac{1}{k!} \sum_{j=i}^N c_j (z_j - t)^k, \quad i \in [N]. \quad (7)$$

Then the statistic in (6) can be succinctly written as

$$T = \max_{i \in [N]} \sup_{t \in [z_{i-1}, z_i]} |\phi_i(t)|, \quad (8)$$

where we let $z_0 = 0$ for convenience. Note each $\phi_i(t)$, $i \in [N]$ is a kth degree polynomial. We can compute a representation for these polynomials efficiently.

Lemma 1. *Fix* $k \geq 0$. *The polynomials in* (7) *satisfy the recurrence relations*

$$\phi_i(t) = \frac{1}{k!} c_i (z_i - t)^k + \phi_{i+1}(t), \quad i \in [N]$$

(where $\phi_{N+1} = 0$*). Given the monomial expansion*

$$\phi_{i+1}(t) = \sum_{\ell=0}^k a_{i+1,\ell} t^\ell,$$

we can compute an expansion for ϕ_i*, with coefficients* $a_{i\ell}$, $\ell \in \{0\} \cup [k]$*, in* $O(1)$ *time. So we can compute all coefficients* $a_{i,\ell}$, $i \in [N]$, $\ell \in \{0\} \cup [k]$ *in* $O(N)$ *time.*

To compute T in (8), we must maximize each polynomial ϕ_i over its domain $[z_{i-1}, z_i]$, for $i \in [N]$, and then compare maxima. Once we have computed a representation for these polynomials, as Lemma 1 ensures we can do in $O(N)$ time, we can use this to analytically maximize each polynomial over its domain, provided the order k is small enough. Of course, maximizing a polynomial over an interval can be reduced to computing the roots of its derivative, which is an analytic computation for any $k \leq 5$ (since the roots of any quartic have a closed-form, see, e.g., Rosen 1995). The next result summarizes.

Proposition 2. *For any* $0 \leq k \leq 5$*, the test statistic in* (8) *can be computed in* $O(N)$ *time.*

Maximizing a polynomial of degree $k \geq 6$ is not generally possible in closed-form. However, developments in semidefinite optimization allow us to approximate its maximum efficiently, investigated next.

2.3 Linear-Time Approximation for $k \geq 6$

Seminal work of Shor (1998); Nesterov (2000) shows that the problem of maximizing a polynomial over an interval can be cast as a semidefinite program (SDP). The number of variables in this SDP depends only on the polynomial order k, and all constraint functions are self-concordant. Using say an interior point method to solve this SDP, therefore, leads to the following result.

Proposition 3. *Fix* $k \geq 6$ *and* $\epsilon > 0$. *For each polynomial in* (7)*, we can compute an* ϵ*-approximation to its maximum in* $c_k \log(1/\epsilon)$ *time, for a constant* $c_k > 0$ *depending only on* k*. As we can compute a representation for all these polynomials in* $O(N)$ *time (Lemma 1), this means we can compute an* ϵ*-approximation to the statistic in* (6) *in* $O(N \log(1/\epsilon))$ *time.*

Remark 3. Let T_ϵ denote the ϵ-approximation from Proposition 3. Under the null $P = Q$, we would need to have $\epsilon = o(1/\sqrt{N})$ in order for the approximation T_ϵ to share the asymptotic null distribution of T, as we will see in Section 3.3. Taking say, $\epsilon = 1/N$, the statistic $T_{1/N}$ requires $O(N \log N)$ computational time, and this is why in various places we make reference to a *nearly* linear-time approximation when $k \geq 6$.

2.4 Simple Linear-Time Approximation

We conclude this section by noting a simple approximation to (6) given by

$$T^* = \max\bigg\{ \max_{t \in Z^0_{(N)},\, t \geq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+|, \\ \max_{t \in Z^0_{(N)},\, t \leq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^-| \bigg\}, \quad (9)$$

where $Z^0_{(N)} = \{0\} \cup Z_{(N)}$. Clearly, for $k = 0$ or 1, the maximizing t in (6) must be one of the sample points $Z_{(N)}$, so $T^* = T$ and there is no approximation error in (9). For $k \geq 2$, we can control the error as follows.

Lemma 2. *For $k \geq 2$, the statistics in (6), (9) satisfy*

$$T - T^* = \frac{\delta_N}{(k-1)!}\bigg(\frac{1}{m}\sum_{i=1}^m |x_i|^{k-1} + \frac{1}{n}\sum_{i=1}^n |y_i|^{k-1} \bigg),$$

where δ_N is the maximum gap between sorted points in $Z^0_{(N)}$.

Remark 4. We would need to have $\delta_N = o_P(1/\sqrt{N})$ in order for T^* to share the asymptotic null of T, see again Section 3.3 (this is assuming that P has $k-1$ moments, so the sample moments concentrate for large enough N). This will not be true of δ_N, the maximum gap, in general. But it does hold when P is continuous, having compact support, and a density bounded from below on its support; here, in fact, $\delta_N = o_P(\log N/N)$ (see, e.g., Wang et al. 2014).

Although it does not have the strong guarantees of the approximation from Proposition 3, the statistic in (9) is simple and efficient—we must emphasize that it can be computed in $O(N)$ linear time, as a consequence of Lemma 1 (the evaluations of $\phi_i(t)$ at the sample points $t \in Z_{(N)}$ are the constant terms a_{i0}, $i \in [N]$ in their monomial expansions)—and is likely a good choice for most practical purposes.

3 ASYMPTOTIC NULL

To study the asymptotic null distribution of the proposed higher-order KS test, we will appeal to uniform central limit theorems (CLTs) from the empirical process theory literature, reviewed here for completeness.

For functions f, g in a class $\mathcal{F}$, let $\mathbb{G}_{P,\mathcal{F}}$ denote a Gaussian process indexed by $\mathcal{F}$ with mean and covariance

$$\mathbb{E}(\mathbb{G}_{P,\mathcal{F}} f) = 0, \quad f \in \mathcal{F},$$
$$\mathrm{Cov}(\mathbb{G}_{P,\mathcal{F}} f,\, \mathbb{G}_{P,\mathcal{F}} g) = \mathrm{Cov}_{X \sim P}(f(X), g(X)), \quad f, g \in \mathcal{F}.$$

For functions l, u, let $[l, u]$ denote the set of functions $\{f : l(x) \leq f(x) \leq u(x), \text{ for all } x\}$. Call $[l, u]$ a *bracket* of size $\|u - l\|_2$, where $\|\cdot\|_2$ denotes the $L_2(P)$ norm, defined as

$$\|f\|_2^2 = \int f(x)^2 \, dP(x).$$

Finally, let $N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{F})$ be the smallest number of ϵ-sized brackets that are required to cover $\mathcal{F}$. Define the bracketing integral of $\mathcal{F}$ as

$$J_{[]}(\|\cdot\|_2, \mathcal{F}) = \int_0^1 \sqrt{\log N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{F})}\, d\epsilon.$$

Note that this is finite when $\log N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{F})$ grows slower than $1/\epsilon^2$. We now state an important uniform CLT from empirical process theory.

Theorem 2 (Theorem 11.1.1 in Dudley 1999)**.** *If $\mathcal{F}$ is a class of functions with finite bracketing integral, then when $P = Q$ and $m, n \to \infty$, the process*

$$\sqrt{\frac{mn}{m+n}}\{\mathbb{P}_m f - \mathbb{Q}_n f\}_{f \in \mathcal{F}}$$

converges weakly to the Gaussian process $\mathbb{G}_{P,\mathcal{F}}$. Hence,

$$\sqrt{\frac{mn}{m+n}} \sup_{f \in \mathcal{F}} |\mathbb{P}_m f - \mathbb{Q}_n f| \xrightarrow{d} \sup_{f \in \mathcal{F}} |\mathbb{G}_{P,\mathcal{F}} f|.$$

3.1 Bracketing Integral Calculation

To derive the asymptotic null of the higher-order KS test, based on its formulation in (5), and Theorem 2, we would need to bound the bracketing integral of $\mathcal{F}_k$. While there are well-known entropy (log covering) number bounds for related function classes (e.g., Birman and Solomyak 1967; Babenko 1979), and the conversion from covering to bracketing numbers is standard, these results unfortunately require the function class to be uniformly bounded in the sup norm, which is certainly not true of $\mathcal{F}_k$.

Note that the representer result in (6) can be written as $T = \rho(P_m, Q_n; \mathcal{G}_k)$, where

$$\mathcal{G}_k = \{g_t^+ : t \geq 0\} \cup \{g_t^- : t \leq 0\}. \quad (10)$$

We can hence instead apply Theorem 2 to $\mathcal{G}_k$, whose bracketing number can be bounded by direct calculation, assuming enough moments on P.

Lemma 3. *Fix $k \geq 0$. Assume $\mathbb{E}_{X \sim P}|X|^{2k+\delta} \leq M < \infty$, for some $\delta > 0$. For the class $\mathcal{G}_k$ in (10), there is a constant $C > 0$ depending only on k, δ such that*

$$\log N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{G}_k) \leq C \log \frac{M^{1 + \frac{\delta(k-1)}{2k+\delta}}}{\epsilon^{2+\delta}}.$$

3.2 Asymptotic Null for Higher-Order KS

Applying Theorem 2 and Lemma 3 to the higher-order KS test statistic (6) leads to the following result.

Theorem 3. *Fix $k \geq 0$. Assume $\mathbb{E}_{X\sim P}|X|^{2k+\delta} < \infty$, for some $\delta > 0$. When $P = Q$, the test statistic in (6) satisfies, as $m, n \to \infty$,*

$$\sqrt{\frac{mn}{m+n}}T \xrightarrow{d} \sup_{g\in\mathcal{G}_k} |\mathbb{G}_{P,k}g|,$$

where $\mathbb{G}_{P,k}$ is an abbreviation for the Gaussian process indexed by the function class $\mathcal{G}_k$ in (10).

Remark 5. When $k = 0$, note that for $t \geq s \geq 0$, the covariance function is

$$\mathrm{Cov}_{X\sim P}(1\{X > s\}, 1\{X > t\}) = F_P(s)(1 - F_P(t)),$$

where F_P denotes the CDF of P. For $s \leq t \leq 0$, the covariance function is again equal to $F_P(s)(1 - F_P(t))$. The supremum of this Gaussian process over $t \in \mathbb{R}$ is that of a Brownian bridge, so Theorem 3 recovers the well-known asymptotic null distribution of the KS test, which (remarkably) does not depend on P.

Remark 6. When $k \geq 1$, it is not clear how strongly the supremum of the Gaussian process from Theorem 3 depends on P; it appears it must depend on the first k moments of P, but is not clear whether it *only* depends on these moments. Section 5 investigates empirically. Currently, we do not have a precise understanding of whether the asymptotic null is useable in practice, and we suggest using a permutation null instead.

3.3 Asymptotic Null Under Approximation

The approximation from Proposition 3 shares the same asymptotic null, provided $\epsilon > 0$ is small enough.

Corollary 1. *Fix $k \geq 0$. Assume $\mathbb{E}_{X\sim P}|X|^{2k+\delta} < \infty$, for some $\delta > 0$. When $P = Q$, as $m, n \to \infty$ such that m/n converges to a positive constant, the test statistic T_ϵ from Proposition 3 converges at a $\sqrt{N}$-rate to the supremum of the same Gaussian process in Theorem 3, provided $\epsilon = o(1/\sqrt{N})$.*

The approximation in (9) shares the same asymptotic null, provided P is continuous with compact support.

Corollary 2. *Fix $k \geq 0$. Assume that P is continuous, compactly supported, with density bounded from below on its support. When $P = Q$, as $m, n \to \infty$ such that m/n converges to a positive constant, the test statistic T^* in (9) converges at a $\sqrt{N}$-rate to the supremum of the same Gaussian process in Theorem 3.*

4 TAIL CONCENTRATION

We examine the convergence of our test statistics to their population analogs. In general, if the population-level IPM $\rho(P, Q; \mathcal{F}_k)$ is large, then the concentration bounds below will imply that the empirical statistic $\rho(P_m, Q_n; \mathcal{F}_k)$ will be large for m, n sufficiently large, and the test will have power.

We first review the necessary machinery, again from empirical process theory. For $p \geq 1$, and a function f of a random variable $X \sim P$, recall the $L_p(P)$ norm is defined as $\|f\|_p = [\mathbb{E}(f(X)^p)]^{1/p}$. For $p > 0$, recall the *exponential Orlicz norm* of order p is defined as

$$\|f\|_{\Psi_p} = \inf\big\{t > 0 : \mathbb{E}[\exp(|X|^p/t^p)] - 1 \leq 1\big\}.$$

(These norms depend on the measure P, since they are defined in terms of expectations with respect to $X \sim P$, though this is not explicit in our notation.)

We now state an important concentration result.

Theorem 4 (Theorems 2.14.2 and 2.14.5 in van der Vaart and Wellner 1996)**.** *Let $\mathcal{F}$ be a class functions with an envelope function F, i.e., $f \leq F$ for all $f \in \mathcal{F}$. Define*

$$W = \sqrt{n} \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f|,$$

and abbreviate $J = J_{[]}(\|\cdot\|, \mathcal{F})$. For $p \geq 2$, if $\|F\|_p < \infty$, then for a constant $c_1 > 0$,

$$[\mathbb{E}(W^p)]^{1/p} \leq c_1\Big(\|F\|_2 J + n^{-1/2+1/p}\|F\|_p\Big),$$

and for $0 < p \leq 1$, if $\|F\|_{\Psi_p} < \infty$, then for a constant $c_2 > 0$,

$$\|W\|_{\Psi_p} \leq c_2\Big(\|F\|_2 J + n^{-1/2}(1 + \log n)^{1/p}\|F\|_{\Psi_p}\Big).$$

The two-sample test statistic $T = \rho(P_m, Q_n; \mathcal{G}_k)$ satisfies (following by a simple argument using convexity)

$$|T - \rho(P, Q; \mathcal{F}_k)| \leq \rho(P, P_m; \mathcal{F}_k) + \rho(Q, Q_n; \mathcal{F}_k).$$

The terms on the right hand side can each be bounded by Theorem 4, where we can use the envelope function $F(x) = |x|^k/k!$ for $\mathcal{G}_k$. Using Markov's inequality, we can then get a tail bound on the statistic.

Theorem 5. *Fix $k \geq 0$. Assume that P, Q both have p moments, where $p \geq 2$ and $p > 2k$. For the statistic in (6), for any $\alpha > 0$, with probability $1 - \alpha$,*

$$|T - \rho(P, Q; \mathcal{G}_k)| \leq c(\alpha)\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right),$$

where $c(\alpha) = c_0\alpha^{-1/p}$, and $c_0 > 0$ is a constant. If P, Q both have finite exponential Orlicz norms of order $0 < p \leq 1$, then the above holds for $c(\alpha) = c_0(\log(1/\alpha))^{1/p}$.

When we assume k moments, the population IPM for $\mathcal{F}_k$ also has a representer in $\mathcal{G}_k$; by Proposition 1, this implies $\rho(\cdot, \cdot; \mathcal{G}_k)$ is also a metric.

Corollary 3. *Fix $k \geq 0$. Assuming P, Q both have k moments, $\rho(P, Q; \mathcal{F}_k) = \rho(P, Q; \mathcal{G}_k)$. Therefore, by Proposition 1, $\rho(\cdot, \cdot; \mathcal{G}_k)$ is a metric (over the space of distributions P, Q with k moments).*

Putting this metric property together with Theorem 5 gives the following.

Corollary 4. *Fix $k \geq 0$. For $\alpha_N = o(1)$ and $1/\alpha_N = o(N^{p/2})$, reject when the higher-order KS test statistic (6) satisfies $T > c(\alpha_N)(1/\sqrt{m} + 1/\sqrt{n})$, where $c(\cdot)$ is as in Theorem 5. For any P, Q that meet the moment conditions of Theorem 5, as $m, n \to \infty$ in such a way that m/n approaches a positive constant, we have type I error tending to 0, and power tending to 1, i.e., the higher-order KS test is* asymptotically powerful.

5 NUMERICAL EXPERIMENTS

We present numerical experiments that examine the convergence of our test statistic to its asymptotic null, its power relative to other general purpose nonparametric tests, and its power when P, Q have densities with local differences. Experiments comparing to the MMD test with a polynomial kernel are deferred to the appendix.

Convergence to Asymptotic Null. In Figure 3, we plot histograms of finite-sample higher-order KS test statistics and their asymptotic null distributions, when $k = 1, 2$. We considered both $P = N(0, 1)$ and $P = \mathrm{Unif}(-\sqrt{3}, \sqrt{3})$ (the uniform distribution standardized to have mean 0 and variance 1). For a total of 1000 repetitions, we drew two sets of samples from P, each of size $m = n = 2000$, then computed the test statistics. For a total of 1000 times, we also approximated the supremum of the Gaussian process from Theorem 3 via discretization. We see that the finite-sample statistics adhere closely to their asymptotic distributions. Interestingly, we also see that the distributions look roughly similar across all four cases considered. Future work will examine more thoroughly.

Comparison to General-Purpose Tests. In Figures 4 and 5, we compare the higher-order KS tests to the KS test, and other widely-used nonparametric tests from the literature: the kernel maximum mean discrepancy (MMD) test (Gretton et al., 2012) with a Gaussian kernel, the energy distance test (Szekely and Rizzo, 2004), and the Anderson-Darling test (Anderson and Darling, 1954). The simulation setup is the same as that in the introduction, where we considered P, Q with different variances, except here we study different means: $P = N(0, 1)$, $Q = N(0.2, 1)$, and different third moments: $P = N(0, 1)$, $Q = t(3)$, where $t(3)$ denotes Student's t-distribution with 3 degrees of freedom. The

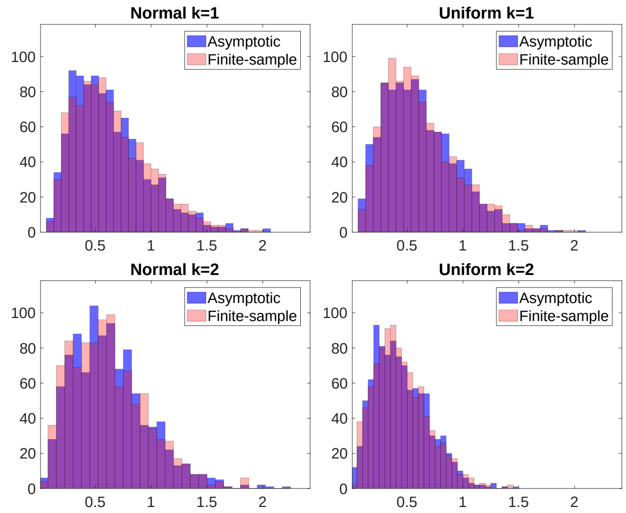


Figure 3: *Histograms comparing finite-sample test statistics to their asymptotic null distribution.*

higher-order KS tests generally perform favorably, and in each setting there is a choice of k that yields better power than KS. In the mean difference setting, this is $k = 1$, and the power degrades for $k = 3, 5$, likely because these tests are "smoothing out" the mean difference too much; see Proposition 4.

Local Density Differences. In Figures 6 and 7, we examine the higher-order KS tests and the KS test, in cases where P, Q have densities p, q such that $p - q$ has sharp local changes. Figure 6 shows a case where $p - q$ is piecewise constant with a few short departures from 0 (see the appendix for a plot) and $m = n = 500$. The KS test is very powerful, and the higher-order KS tests all perform poorly; in fact, the KS test here has better power than all commonly-used nonparametric tests we tried (results not shown). Figure 7 displays a case where $p - q$ changes sharply in the right tail (see the appendix for a plot) and $m = n = 2000$. The power of the higher-order KS test appears to increase with k, likely because the witness functions are able to better concentrate on sharp departures for large k.

6 DISCUSSION

This paper began by noting the variational characterization of the classical KS test as an IPM with respect to functions of bounded total variation, and then proposed a generalization to higher-order total variation classes. This generalization was nontrivial, with subtleties arising in defining the right class of functions so that the statistic was finite and amenable for simplification via a representer result, challenges in computing the statistic efficiently, and challenges in studying asymptotic convergence and concentration due to the fact that the

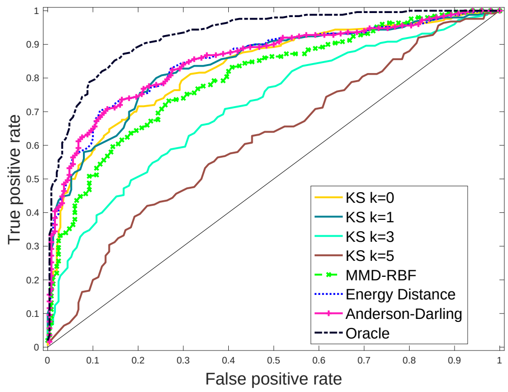


Figure 4: *ROC curves for* $P = N(0, 1)$, $Q = N(0.2, 1)$.

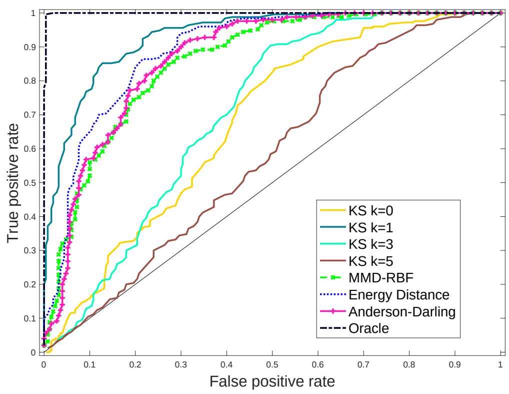


Figure 5: *ROC curves for* $P = N(0, 1)$, $Q = t(3)$.

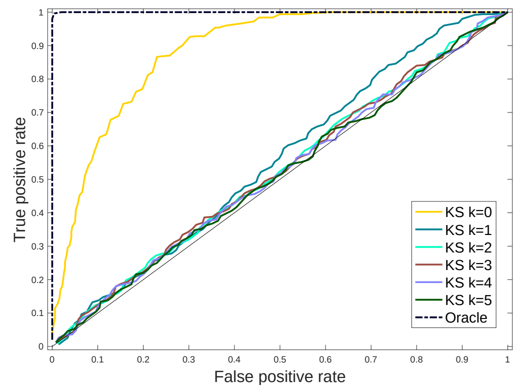


Figure 6: *ROC curves for piecewise constant* $p - q$.

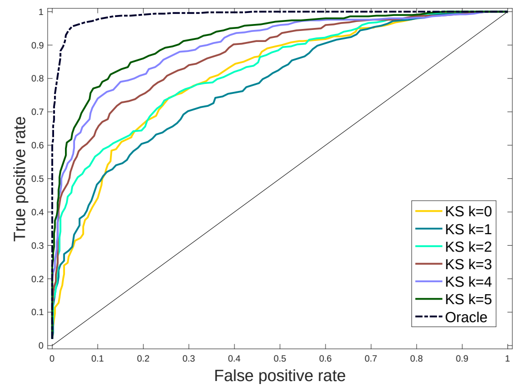


Figure 7: *ROC curves for tail departure in* $p - q$.

function class is not uniformly sup norm bounded. The resulting class of linear-time higher-order KS tests was shown empirically to be more sensitive to tail differences than the usual KS test, and to have competitive power relative to several other popular tests.

In future work, we intend to more formally study the power properties of our new higher-order tests relative to the KS test. The following is a lead in that direction. For $k \geq 1$, define I^k to be the kth order integral operator, acting on a function f, via

$$(I^k f)(x) = \int_0^x \int_0^{t_k} \cdots \int_0^{t_2} f(t_1)\, dt_1 \cdots dt_k.$$

Denote by F_P, F_Q the CDFs of the distributions P, Q. Notice that the population-level KS test statistic can be written as $\rho(P, Q; \mathcal{F}_0) = \|F_P - F_Q\|_\infty$, where $\|\cdot\|_\infty$ is the sup norm. Interestingly, a similar representation holds for the higher-order KS tests.

Proposition 4. *Assuming* P, Q *have* k *moments,*

$$\rho(P, Q; \mathcal{F}_k) = \|(I^k)^*(F_P - F_Q)\|_\infty,$$

where $(I^k)^*$ *is the adjoint of the bounded linear operator* I^k*, with respect to the usual* L_2 *inner product. Further, if* P, Q *are supported on* $[0, \infty)$*, or their first* k *moments match, then we have the more explicit representation*

$$\rho(P, Q; \mathcal{F}_k) = \sup_{x \in \mathbb{R}} \left| \int_x^\infty \int_{t_k}^\infty \cdots \int_{t_2}^\infty (F_P - F_Q)(t_1)\, dt_1 \cdots dt_k \right|.$$

The representation in Proposition 4 could provide one avenue for power analysis. When P, Q are supported on $[0, \infty)$, or have k matching moments, the representation is particularly simple in form. This form confirms the intuition that detecting higher-order moment differences is hard: as k increases, the k-times integrated CDF difference $F_P - F_Q$ becomes smoother, and hence the differences are less accentuated.

In future work, we also intend to further examine the asymptotic null of the higher-order KS test (the Gaussian process from Theorem 3), and determine to what extent it depends on the underlying distribution P (beyond say, its first k moments). Lastly, some ideas in this paper seem extendable to the multivariate and graph settings, another direction for future work.

Acknowledgments. We thank Alex Smola for several early inspiring discussions. VS and RT were supported by NSF Grant DMS-1554123.

A Appendix

A.1 Comparing the Test in Wang et al. (2014)

The test statistic in Wang et al. (2014) can be expressed as

$$T^{**} = \max_{t \in Z_{(N)}} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+| = \max_{t \in Z_{(N)}} \left| \frac{1}{m} \sum_{i=1}^m (x_i - t)_+^k - \frac{1}{n} \sum_{i=1}^n (y_i - t)_+^k \right|. \tag{11}$$

This is very close to our approximate statistic T^* in (9). The only difference is that we replace $g_t^+(x) = (x-t)_+^k$ by $g_t^-(x) = (t-x)_+^k$ for $t \le 0$.

Our exact (not approximate) statistic is in (6). This has the advantage having an equivalent variational form (5), and the latter form is important because it shows the statistic to be a metric.

A.2 Proof of Proposition 1

We first claim that $F(x) = |x|^k/k!$ is an envelope function for $\mathcal{F}_k$, meaning $f \le F$ for all $f \in \mathcal{F}_k$. To see this, note each $f \in \mathcal{F}_k$ has kth weak derivative with left or right limit of 0 at 0, so $|f^{(k)}(x)| \le \mathrm{TV}(f^{(k)}) \le 1$; repeatedly integrating and applying the derivative constraints yields the claim. Now due to the envelope function, if P, Q have k moments, then the IPM is well-defined: $|\mathbb{P}f| < \infty$, $|\mathbb{Q}f| < \infty$ for all $f \in \mathcal{F}_k$. Thus if $P = Q$, then clearly $\rho(P, Q; \mathcal{F}_k) = 0$.

For the other direction, suppose that $\rho(P, Q; \mathcal{F}_k) = 0$. By simple rescaling, for any f, if $\mathrm{TV}(f^{(k)}) = R > 0$, then $\mathrm{TV}((f/R)^{(k)}) \le 1$. Therefore $\rho(P, Q; \mathcal{F}_k) = 0$ implies $\rho(P, Q; \widetilde{\mathcal{F}}_k) = 0$, where

$$\widetilde{\mathcal{F}}_k = \big\{ f : \mathrm{TV}(f^{(k)}) < \infty,\ f^{(j)}(0) = 0,\ j \in \{0\} \cup [k-1],\ \text{and}\ f^{(k)}(0+) = 0 \text{ or } f^{(k)}(0-) = 0 \big\}.$$

This also implies $\rho(P, Q; \widetilde{\mathcal{F}}_k^+) = 0$, where

$$\widetilde{\mathcal{F}}_k^+ = \{ f : \mathrm{TV}(f^{(k)}) < \infty,\ f(x) = 0 \text{ for } x \le 0 \}.$$

As the class $\widetilde{\mathcal{F}}_k^+$ contains $C_c^\infty(\mathbb{R}_+)$, where $\mathbb{R}_+ = \{x : x > 0\}$ (and $C_c^\infty(\mathbb{R}_+)$ is the class of infinitely differentiable, compactly supported functions on $\mathbb{R}_+$), we have by Lemma 4 that $P(A \cap \mathbb{R}_+) = Q(A \cap \mathbb{R}_+)$ for all open sets A. By similar arguments, we also get that $P(A \cap \mathbb{R}_-) = Q(A \cap \mathbb{R}_-)$, for all open sets A, where $\mathbb{R}_- = \{x : x < 0\}$. This implies that $P(\{0\}) = Q(\{0\})$ (as $1 - P(\mathbb{R}_+) - P(\mathbb{R}_-)$, and the same for Q), and finally, $P(A) = Q(A)$ for all open sets A, which means that $P = Q$.

A.3 Statement and Proof of Lemma 4

Lemma 4. *For any two distributions P, Q supported on an open set Ω, if $\mathbb{E}_{X \sim P}[f(X)] = \mathbb{E}_{Y \sim Q}[f(Y)]$ for all $f \in C_c^\infty(\Omega)$, then $P = Q$.*

Proof. It suffices to show that $P(A) = Q(A)$ for every open set $A \subseteq \Omega$. As P, Q are probability measures and hence Radon measures, there exists a sequence of compact sets $K_n \subseteq A$, $n = 1, 2, 3, \ldots$ such that $\lim_{n\to\infty} P(K_n) = P(A)$ and $\lim_{n\to\infty} Q(K_n) = Q(A)$. Let f_n, $n = 1, 2, 3, \ldots$ be smooth compactly supported functions with values in $[0, 1]$ such that $f_n = 1$ on K_n and $f_n = 0$ outside of A. (Such functions can be obtained by applying Urysohn's Lemma on appropriate sets containing K_n and A and convolving the resulting continuous function with a bump function.) Then $P(K_n) \le E_P(f_n) = E_Q(f_n) \le Q(A)$ (where the equality by the main assumption in the lemma). Taking $n \to \infty$ gives $P(A) \le Q(A)$. By reversing the roles of P, Q, we also get $Q(A) \le P(A)$. Thus $P(A) = Q(A)$. ☐

A.4 Proof of Theorem 1

Let $\mathcal{G}_k$ be as in (10). Noting that $G_k \subseteq \mathcal{F}_k$, it is sufficient to show

$$\sup_{f \in \mathcal{F}_k} |\mathbb{P}_m f - \mathbb{Q}_n f| \le \sup_{g \in \mathcal{G}_k} |\mathbb{P}_m g - \mathbb{Q}_n g|.$$

Fix any $f \in \mathcal{F}_k$. Denote $Z^0_{(N)} = \{0\} \cup Z_{(N)}$. From the statement and proof of Theorem 1 in Mammen (1991), there exists a spline $\tilde{f}$ of degree k, with finite number of knots such that for all $z \in Z^0_{(N)}$

$$\begin{aligned} f(z) &= \tilde{f}(z), \\ f^{(j)}(z) &= \tilde{f}^{(j)}(z), \quad j \in [k-1], \\ f^{(k)}(z^+) &= \tilde{f}^{(k)}(z^+), \\ f^{(k)}(z^-) &= \tilde{f}^{(k)}(z^-). \end{aligned}$$

and importantly, $\mathrm{TV}(\tilde{f}^{(k)}) \le \mathrm{TV}(f^{(k)})$. As $0 \in Z^0_{(N)}$, we hence know that the boundary constraints (derivative conditions at 0) are met, and $\tilde{f} \in \mathcal{F}_k$.

Because $\tilde{f}$ is a spline with a given finite number of knot points, we know that it has an expansion in terms of truncated power functions. Write $t_0, t_1, \ldots, t_L$ for the knots of $\tilde{f}$, where $t_0 = 0$. Also denote $g_t = g_t^+$ when $t > 0$, and $g_t = g_t^-$ when $t < 0$. Then for some $\alpha_\ell \in \mathbb{R}$, $\ell \in \{0\} \cup [L]$, and a polynomial p of degree k, we have

$$\tilde{f} = p + \alpha_0 g_0^+ + \sum_{\ell=1}^{L} \alpha_\ell g_{t_\ell},$$

The boundary conditions on $\tilde{f}$, g_0^+, g_{t_ℓ}, $\ell \in [L]$ imply

$$\begin{gathered} p(0) = p^{(1)}(0) = \ldots = p^{(k-1)}(0) = 0, \\ (\alpha_0 g_{0^+} + p)^{(k)}(0^+) = 0 \;\text{ or }\; (\alpha_0 g_{0^+} + p)^{(k)}(0^-) = 0. \end{gathered}$$

The second line above implies that

$$\alpha_0 + p^{(k)} = 0 \;\text{ or }\; p^{(k)} = 0.$$

In the second case, we have $p = 0$. In the first case, we have $p(x) = -\alpha_0 x^k / k!$, so $\alpha_0 g_0 + p = -(-1)^{k+1} \alpha_0 g_0^-$. Therefore, in all cases we can write

$$\tilde{f} = \sum_{\ell=0}^{L} \alpha_\ell g_{t_\ell},$$

with the new understanding that g_0 is either g_0^+ or g_0^-. This means that $\tilde{f}$ lies in the span of functions in $\mathcal{G}_k$. Furthermore, our last expression for $\tilde{f}$ implies

$$\|\alpha\|_1 = \sum_{\ell=0}^{L} |\alpha_\ell| = \mathrm{TV}(\tilde{f}^{(k)}) \le \mathrm{TV}(f^{(k)}) \le 1.$$

Finally, using the fact that f and $\tilde{f}$ agree on $Z^0_{(N)}$,

$$\begin{aligned} |\mathbb{P}_m f - \mathbb{Q}_n f| &= |\mathbb{P}_m \tilde{f} - \mathbb{Q}_n \tilde{f}| \\ &= \left| \sum_{\ell=0}^{L} \alpha_\ell (\mathbb{P}_m g_{t_\ell} - \mathbb{Q}_n g_{t_\ell}) \right| \\ &\le \sum_{\ell=0}^{L} |\alpha_\ell| \cdot \sup_{g \in \mathcal{G}_k} |\mathbb{P}_m g - \mathbb{Q}_n g| \\ &\le \sup_{g \in \mathcal{G}_k} |\mathbb{P}_m g - \mathbb{Q}_n g|, \end{aligned}$$

the last two lines following from Holder's inequality, and $\|\alpha\|_1 \le 1$. This completes the proof.

A.5 Proof of Proposition 3

From Shor (1998); Nesterov (2000), a polynomial of degree $2d$ is nonnegative on $\mathbb{R}$ if and only if it can be written as a sum of squares (SOS) of polynomials, each of degree d. Crucially, one can show that $p(x) = \sum_{i=0}^{2d} a_i x^i$ is SOS if and only if there is a positive semidefinite matrix $Q \in \mathbb{R}^{(d+1)\times(d+1)}$ such that

$$a_{i-1} = \sum_{j+k=i} Q_{jk}, \quad i \in [2d].$$

Finding such a matrix Q can be cast as a semidefinite program (SDP) (a feasibility program, to be precise), and therefore checking nonnegativity can be done by solving an SDP.

Furthermore, calculating the maximum of a polynomial p is equivalent to calculating the smallest γ such that $\gamma - p$ is nonnegative. This is therefore also an SDP.

Finally, a polynomial of degree k is nonnegative an interval $[a, b]$ if and only if it can be written as

$$p(x) = \begin{cases} s(x) + (x-a)(b-x)t(x) & k \text{ even} \\ (x-a)s(x) + (b-x)t(x) & k \text{ odd} \end{cases}, \tag{12}$$

where s, t are polynomials that are both SOS. Thus maximizing a polynomial over an interval is again equivalent to an SDP. For details, including a statement that such an SDP can be solved to ϵ-suboptimality in $c_k \log(1/\epsilon)$ iterations, where $c_k > 0$ is a constant that depends on k, see Nesterov (2000).

A.6 Proof of Lemma 2

Suppose t^* maximizes the criterion in (6). If $t^* = 0$, then $T^* = T$ and the result trivially holds. Assume without a loss of generality that $t^* > 0$, as the result for $t^* < 0$ will follow similarly.

If t^* is one of the sample points $Z_{(N)}$, then $T^* = T$ and the result trivially holds; if t^* is larger than all points in $Z_{(N)}$, then $T^* = T = 0$ and again the result trivially holds. Hence we can assume without a loss of generality that $t^* \in (a, b)$, where $a, b \in Z^0_{(N)}$. Define

$$\phi(t) = \frac{1}{k!} \sum_{i=1}^N c_i (z_i - t)_+^k, \;\; t \in [a, b],$$

where $c_i = (\mathbb{1}_m/m - \mathbb{1}_n/n)_i$, $i \in [N]$, as before. Note that $T = \phi(t^*)$, and

$$|\phi'(t)| \leq \frac{1}{(k-1)!} \sum_{i=1}^N |c_i| |z_i^{k-1}| = \frac{1}{(k-1)!} \left(\frac{1}{m} \sum_{i=1}^m |x_i|^{k-1} + \frac{1}{n} \sum_{i=1}^n |y_i|^{k-1} \right) := L.$$

Therefore

$$T - T^* \leq |f(t^*)| - |f(a)| \leq |f(t^*) - f(a)| \leq |t^* - a| L \leq \delta_N L,$$

as desired.

A.7 Proof of Lemma 3

Decompose $\mathcal{G}_k = \mathcal{G}_k^+ \cup \mathcal{G}_k^-$, where $\mathcal{G}_k^+ = \{g_t^+ : t \geq 0\}$, $\mathcal{G}_k^- = \{g_t^- : t \leq 0\}$. We will bound the bracketing number of $\mathcal{G}_k^+$, and the result for $\mathcal{G}_k^-$, and hence $\mathcal{G}_k$, follows similarly.

Our brackets for $\mathcal{G}_k^+$ will be of the form $[g_{t_i}, g_{t_{i+1}}]$, $i \in \{0\} \cup [R]$, where $0 = t_1 < t_2 < \cdots < t_{R+1} = \infty$ are to be specified, with the convention that $g_\infty = 0$. It is clear that such a set of brackets covers $\mathcal{G}_k^+$. Given $\epsilon > 0$, we need to choose the brackets such that

$$\|g_{t_i} - g_{t_{i+1}}\|_2 \leq \epsilon, \;\; i \in \{0\} \cup [R], \tag{13}$$

and then show that the number of brackets R is small enough to satisfy the bound in the statement of the lemma.

For any $0 \leq s < t$,

$$\begin{aligned} k!^2 \|g_s - g_t\|_2^2 &= \int_s^t (x-s)_+^{2k} \, dP(x) + \int_t^\infty \big((x-s)^k - (x-t)^k \big)^2 \, dP(x) \\ &\leq \int_s^\infty \big(k(x-s)^{k-1}(t-s) \big)^2 \, dP(x) \\ &= k^2 (t-s)^2 \int_s^\infty (x-s)^{2k-2} \, dP(x), \end{aligned}$$

where the second line follows from elementary algebra. Now in view of the moment bound assumption, we can bound the integral above using Holder's inequality with $p = (2k+\delta)/(2k-2)$ and $q = (2k+\delta)/(2+\delta)$ to get

$$\begin{aligned} k!^2 \|g_s - g_t\|_2^2 &\leq k^2(t-s)^2 \left(\int_s^\infty (x-s)^{2k+\delta} \, dP(x) \right)^{1/p} \left(\int_s^\infty 1^q \, (x) dP \right)^{1/q} \\ &\leq \frac{M^{1/p}}{(k-1)!^2} (t-s)^2, \end{aligned} \tag{14}$$

where recall the notation $M = \mathbb{E}[|X|^{2k+\delta}] < \infty$.

Also, for any $t > 0$, using Holder's inequality again, we have

$$\begin{aligned} k!^2 \|g_t - 0\|_2^2 &= \int_t^\infty (x-t)^{2k} \, dP(x) \\ &\leq \left(\int_t^\infty (x-t)^{2k+\delta} \, dP(x) \right)^{2k/(2k+\delta)} \big(P(X \geq t) \big)^{\delta/(2k+\delta)} \\ &\leq M^{2k/(2k+\delta)} \left(\frac{\mathbb{E}|X|^{2k+\delta)}}{t^{2k+\delta}} \right)^{\delta/(2k+\delta)} = \frac{M}{t^\delta}, \end{aligned} \tag{15}$$

where in the third line we used Markov's inequality.

Fix an $\epsilon > 0$. For parameters $\beta, R > 0$ to be determined, set $t_i = (i-1)\beta$ for $i \in [R]$ and $t_0 = 0$, $t_{R+1} = \infty$. Looking at (14), to meet (13), we see we can choose β such that

$$\frac{M^{1/p}}{(k-1)!^2} \beta^2 \leq \epsilon^2.$$

Then for such a β, looking at (15), we see we can choose R such that

$$\frac{M}{k!^2((R-1)\beta)^\delta} \leq \epsilon^2.$$

In other words, we can choose choose

$$\beta = \frac{(k-1)!}{M^{1/2p}}, \quad R = 1 + \left\lceil \frac{M^{1/2p+1/\delta}}{(k-1)! k!^{2/\delta} \epsilon^{2/\delta+1}} \right\rceil,$$

and (14), (15) imply that we have met (13). Therefore,

$$\log N_{[]}(\epsilon, \|\cdot\|, \mathcal{G}_k^+) \leq \log R \leq C \log \frac{M^{1+\frac{\delta(k-1)}{2k+\delta}}}{\epsilon^{2+\delta}},$$

where $C > 0$ depends only on k, δ.

A.8 Proof of Theorem 3

Once we have a finite bracketing integral for $\mathcal{G}_k$, we can simply apply Theorem 2 to get the result. Lemma 3 shows the log bracketing number of $\mathcal{G}_k$ to grow at the rate $\log(1/\epsilon)$, slow enough to imply a finite bracketing integral (the bracketing integral will be finite as long as the log bracketing number does not grow faster than $1/\epsilon^2$).

A.9 Proof of Corollaries 1 and 2

For the approximation from Proposition 3, observe

$$\sqrt{N} T_\epsilon = \sqrt{N} T + \sqrt{N}(T - T_\epsilon),$$

and $0 \leq \sqrt{N}(T - T_\epsilon) \leq \sqrt{N}\epsilon$, so for $\epsilon = o(1/\sqrt{N})$, we will have $\sqrt{N} T_\epsilon$ converging weakly to the same Gaussian process as $\sqrt{N} T$.

For the approximation in (9), the argument is similar, and we are simply invoking Lemma 5 in Wang et al. (2014) to bound the maximum gap δ_N in probability, under the density conditions.

A.10 Proof of Theorem 5

Let $W = \sqrt{m}\rho(P_m, P; \mathcal{G}_k)$. The bracketing integral of $\mathcal{G}_k$ is finite due to the slow growth of the log bracketing number from Lemma 3, at the rate $\log(1/\epsilon)$. Also, we can clearly take $F(x) = |x|^k/k!$ as an envelope function for $\mathcal{G}_k$. Thus, we can apply Theorem 5 to yield

$$\left(\mathbb{E}[\rho(P_m, P; \mathcal{G}_k)^p]\right)^{1/p} \leq \frac{C}{\sqrt{m}}$$

for a constant $C > 0$ depending only on k, p, and $\mathbb{E}|X|^p$. Combining this with Markov's inequality, for any a,

$$\mathbb{P}\big(\rho(P_m, P; \mathcal{G}_k) > a\big) \leq \left(\frac{C}{\sqrt{m}a}\right)^p,$$

thus for $a = C/(\sqrt{m}\alpha^{1/p})$, we have $\rho(P_m, P; \mathcal{G}_k) \leq a$ with probability at least $1 - \alpha$. The same argument applies to $W = \sqrt{n}\rho(Q_n, P; \mathcal{G}_k)$, and putting these together yields the result. The result when we additionally assume finite Orlicz norms is also similar.

A.11 Proof of Corollary 3

Let f maximize $|(\mathbb{P} - \mathbb{Q})f|$. Due to the moment conditions (see the proof of Proposition 1), we have $|\mathbb{P}f| < \infty$, $|\mathbb{Q}f| < \infty$. Assume without loss of generality that $(\mathbb{P} - \mathbb{Q})f > 0$. By the strong law of large numbers, we have $(\mathbb{P}_m - \mathbb{Q}_n)f \to (\mathbb{P} - \mathbb{Q})f$ as $m, n \to \infty$, almost surely. Also by the strong law, $\mathbb{P}_m|x|^{k-1} \to \mathbb{P}|x|^{k-1}$ as $m \to \infty$, almost surely, and $\mathbb{Q}_n|y|^{k-1} \to \mathbb{Q}|y|^{k-1}$ as $n \to \infty$, almost surely. For what follows, fix any samples $X_{(m)}, Y_{(n)}$ (i.e., take them to be nonrandom) such that the aforementioned convergences hold.

For each m, n, we know by the representer result in Theorem 1 that there exists $g_{mn} \in \mathcal{G}_k$ such that $(\mathbb{P}_m - \mathbb{Q}_n)f \leq |(\mathbb{P}_m - \mathbb{Q}_n)g_{mn}|$. (This is possible since the proof of Theorem 1 does not rely on any randomness that is inherent to $X_{(m)}, Y_{(n)}$, and indeed it holds for any fixed sets of samples.) Assume again without a loss of generality that $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} > 0$. Denote by t_{mn} the knot of g_{mn} (i.e., $g_{mn} = g^+_{t_{mn}}$ if $t \geq 0$, and $g_{mn} = g^-_{t_{mn}}$ if $t \leq 0$). We now consider two cases.

If $|t_{mn}|$ is a bounded sequence, then by the Bolzano-Weierstrass theorem, it has a convergent subsequence, which converges say to $t \geq 0$. Passing to this subsequence (but keeping the notation unchanged, to avoid unnecessary clutter) we claim that $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} \to (\mathbb{P} - \mathbb{Q})g$ as $m, n \to \infty$, where $g = g^+_t$. To see this, assume $t_{mn} \geq t$ without a loss of generality (the arguments for $t_{mn} \leq t$ are similar), and note

$$g(x) - g_{mn}(x) = \frac{1}{k!}\begin{cases} 0 & x < t \\ (x-t)^k & t \leq x < t_{mn} \\ (t_{mn} - t)\sum_{i=0}^{k-1}(x-t)^i(x-t_{mn})^{k-1-i} & x \geq t_{mn} \end{cases},$$

where we have used the identity $a^k - b^k = (a-b)\sum_{i=0}^{k-1} a^i b^{k-1-i}$. Therefore, as $m, n \to \infty$,

$$|\mathbb{P}_m(g_{mn} - g)| \leq |t_{mn} - t| \, \mathbb{P}_m|x|^{k-1}/(k-1)! \to 0,$$

because $t_{mn} \to t$ by definition, and $\mathbb{P}_m|x|^{k-1} \to \mathbb{P}|x|^{k-1}$. Similarly, as $m, n \to \infty$, we have $|\mathbb{Q}_n(g_{mn} - g)| \to 0$, and therefore $|(\mathbb{P}_m - \mathbb{Q}_n)(g_{mn} - g)| \leq |\mathbb{P}_m(g_{mn} - g)| + |\mathbb{Q}_m(g_{mn} - g)| \to 0$, which proves the claim. But since $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} \geq (\mathbb{P}_m - \mathbb{Q}_n)f$ for each m, n, we must have $(\mathbb{P} - \mathbb{Q})g \geq (\mathbb{P} - \mathbb{Q})f$ and hence $(\mathbb{P} - \mathbb{Q})g = (\mathbb{P} - \mathbb{Q})f$ by definition of f because $g \in \mathcal{F}_k$, i.e., there is a representer in $\mathcal{G}_k$, as desired.

If $|t_{mn}|$ is unbounded, then pass to a subsequence in which t_{mn} converges say to ∞ (the case for convergence to $-\infty$ is similar). In this case, we have $(\mathbb{P}_m - Q_n)g_{mn} \to 0$ as $m, n \to \infty$, and since $|(\mathbb{P}_m - \mathbb{Q}_n)g_{mn}| \geq (\mathbb{P}_m - \mathbb{Q}_n)f$ for each m, n, we have $(\mathbb{P} - \mathbb{Q})f = 0$. But we can achieve this with $(\mathbb{P} - \mathbb{Q})g^+_t$, by taking $t \to \infty$, so again we have a representer in $\mathcal{G}_k$, as desired.

A.12 Proof of Corollary 4

When we reject as specified in the corollary, note that for $P = Q$, we have type I error at most α_N by Theorem 4, and as $\alpha_N = o(1)$, we have type I error converging to 0.

For $P \neq Q$, such that the moment conditions are met, we know by Corollary 3 that $\rho(P, Q; \mathcal{G}_k) \neq 0$. Recalling $1/\alpha_N = o(N^{p/2})$, we have as $N \to \infty$,

$$c(\alpha_N)\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) = \alpha^{-1/p}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) \to 0.$$

The concentration result from Theorem 5 shows that T will concentrate around $\rho(P, Q; \mathcal{G}_k) \neq 0$ with probability tending to 1, and thus we reject with probability tending to 1.

A.13 Additional Experiments

A.14 Local Density Differences Continued

Figure 8 plots the densities used for the local density difference experiments, with the left panel corresponding to Figure 6, and the right panel to Figure 7.

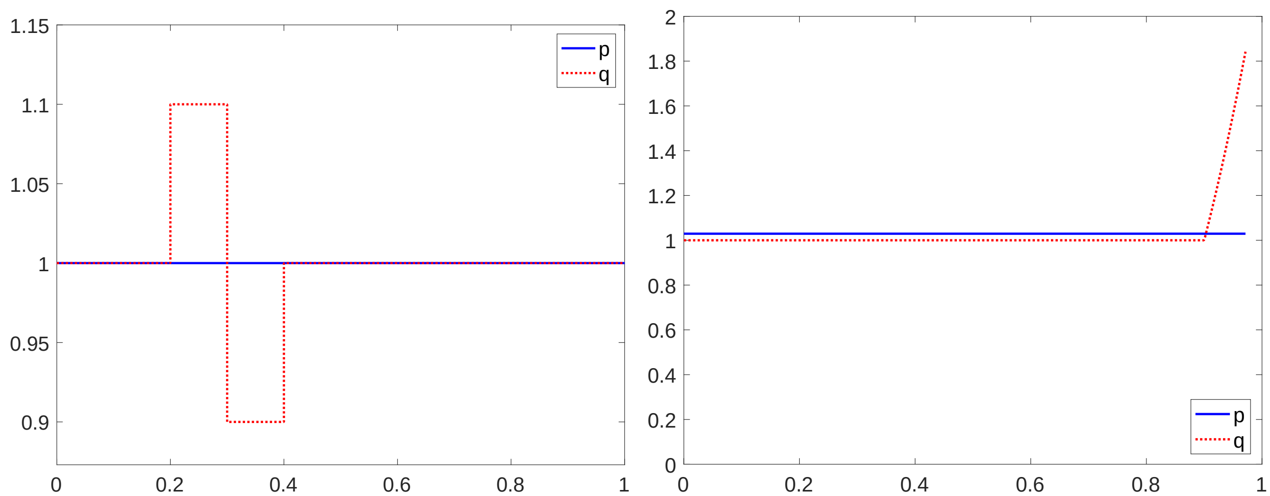


Figure 8: *Densities for the local density difference experiments.*

A.15 Comparison to MMD with Polynomial Kernel

Now we compare the higher-order KS test to the MMD test with a polynomial kernel, as suggested by a referee of this paper. The MMD test with a polynomial kernel looks at moment differences up to some prespecified order $d \geq 1$, and its test statistic can be written as

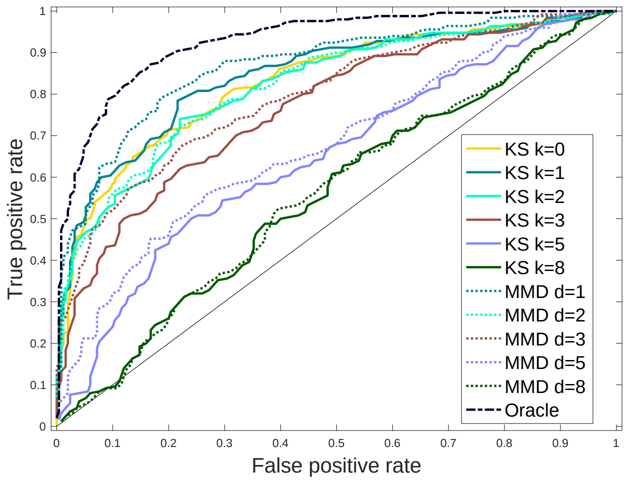
$$\sum_{i=0}^{d} \binom{d}{i} (\mathbb{P}_n x^i - \mathbb{P}_m y^i)^2.$$

This looks at a weighted sum of *all* moments up to order d, whereas our higher-order KS test looks at truncated moments of a *single* order k. Therefore, to put the methods on more equal footing, we aggregated the higher-order KS test statistics up to order k, i.e., writing T_i to denote the ith order KS test statistic, $i \in [k]$, we considered

$$\sum_{i=0}^{k} \binom{k}{i} T_i^2,$$

borrowing the choice of weights from the MMD polynomial kernel test statistic.

Figure 9 shows ROC curves from two experiments comparing the higher-order KS test and MMD polynomial kernel tests. We used distributions $P = N(0, 1)$, $Q = N(0.2, 1)$ in the left panel (as in Figure 4), and $P = N(0, 1)$, $Q = t(3)$ in the right panel (as in Figure 5). We can see that the (aggregated) higher-order KS tests and MMD polynomial kernel tests perform roughly similarly.



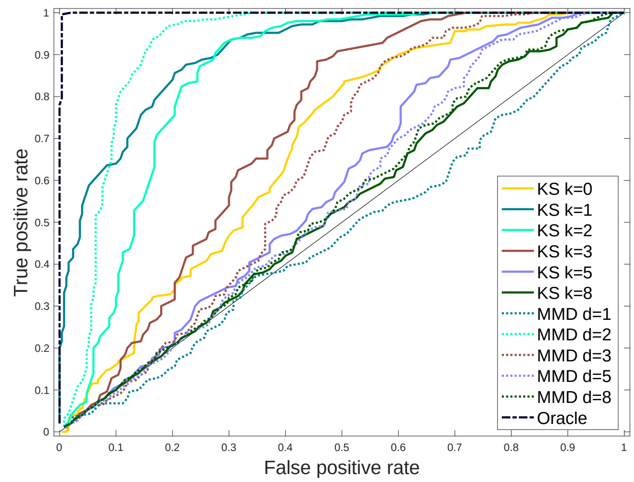


Figure 9: *ROC curves for* $P = N(0, 1)$, $Q = N(0.2, 1)$ *(left), and* $P = N(0, 1)$, $Q = t(3)$ *(right).*

There is one important point to make clear: the population MMD test with a polynomial kernel is *not* a metric, i.e., there are distributions $P \neq Q$ for which the population-level test statistic is exactly 0. This is because it only considers moment differences up to order d, thus any pair of distributions P, Q that match in the first d moments but differ in (say) the $(d+1)$st will lead to a population-level statistic that 0. In this sense, the MMD test with a polynomial kernel is not truly nonparametric, whereas the KS test, the higher-order KS tests the MMD test with a Gaussian kernel, the energy distance test, the Anderson-Darling test, etc., all are.

A.16 Proof of Proposition 4

For $k \geq 1$, recall our definition of I^k the kth order integral operator,

$$(I^k f)(x) = \int_0^x \int_0^{t_k} \cdots \int_0^{t_2} f(t_1)\, dt_1 \cdots dt_k,$$

Further, for $k \geq 1$, denote by D^k the kth order derivative operator,

$$(D^k f)(x) = f^{(k)}(x),$$

Is it not hard to check that over all functions f with k weak derivatives, and that obey the boundary conditions $f(0) = f'(0) = \cdots = f^{(k-1)}(0) = 0$, these two operators act as inverses, in that

$$D^k I^k f = f, \ \text{ and } \ I^k D^k f = f.$$

For a measure μ, denote $\langle f, d\mu \rangle = \int f(x)\, d\mu(x)$. (This is somewhat of an abuse of the notation for the usual L_2 inner product on square integrable functions, but it is convenient for what follows.) With this notation, we can write the kth order KS test statistic, at the population-level, as

$$\begin{aligned} \sup_{f \in \mathcal{F}_k} |\mathbb{P}f - \mathbb{Q}f| &= \sup_{f \in \mathcal{F}_k} |\langle f, dP - dQ \rangle| \\ &= \sup_{f \in \mathcal{F}_k} |\langle I^k D^k f, dP - dQ \rangle| \\ &= \sup_{\substack{h: \mathrm{TV}(h) \leq 1, \\ h(0+)=0 \text{ or } h(0-)=0}} |\langle I^k h, dP - dQ \rangle| \\ &= \sup_{\substack{h: \mathrm{TV}(h) \leq 1, \\ h(0+)=0 \text{ or } h(0-)=0}} |\langle h, (I^k)^*(dP - dQ) \rangle| \\ &= \|(I^1)^*(I^k)^*(dP - dQ)\|_\infty. \end{aligned} \tag{16}$$

In the second line, we used the fact that I^k and D^k act as inverses over $f \in \mathcal{F}_k$ because these functions all satisfy the appropriate boundary conditions. In the third line, we simply reparametrized via $h = f^{(k)}$. In the fourth line, we introduced the adjoint operator $(I^k)^*$ of I^k (which will be described in detail shortly). In the fifth line, we leveraged the variational result for the KS test ($k = 0$ case), where $(I^1)^*$ denotes the adjoint of the integral operator I^1 (details below), and we note that the limit condition at 0 does do not affect the result here.

We will now study the adjoints corresponding to the integral operators. By definition $(I^1)^* g$ must satisfy for all functions f

$$\langle I^1 f, g\rangle = \langle f, (I^1)^* g\rangle.$$

We can rewrite this as

$$\int \int_0^x f(t) g(x)\, dt\, dx = \int f(t)((I^1)^* g)(t)\, dt,$$

and we can recognize by Fubini's theorem that therefore

$$((I^1)^* g)(t) = \begin{cases} \displaystyle\int_t^\infty g(x)\, dx & t \geq 0 \\ \displaystyle -\int_{-\infty}^t g(x)\, dx & t < 0. \end{cases}$$

For functions g that integrate to 0, this simplifies to

$$((I^1)^* g)(t) = \int_t^\infty g(x) dx, \quad t \in \mathbb{R}. \tag{17}$$

Returning to (16), because we can decompose $I^k = I^1 I^1 \cdots I^1$ (k times composition), it follows that $(I^k)^* = (I^1)^*(I^1)^* \cdots (I^1)^*$ (k times composition), so

$$\|(I^1)^*(I^k)^*(dP - dQ)\|_\infty = \|(I^k)^*(I^1)^*(dP - dQ)\|_\infty = \|(I^k)^*(F_P - F_Q)\|_\infty,$$

where in the last step we used (17), as $dP - dQ$ integrates to 0. This proves the first result in the proposition.

To prove the second result, we will show that

$$(I^k)^*(F_P - F_Q)(x) = \int_x^\infty \int_{t_k}^\infty \cdots \int_{t_2}^\infty (F_P - F_Q)(t_1)\, dt_1 \cdots dt_k,$$

when P, Q has nonnegative supports, or have k matching moments. In the first case, the above representation is clear from the definition of the adjoint. In the second case, we proceed by induction on k. For $k = 1$, note that $F_P - F_Q$ integrates to 0, which is true because

$$\langle 1, F_P - F_Q\rangle = \langle 1, (I^1)^*(dP - dQ)\rangle = \langle x, dP - dQ\rangle = 0,$$

the last step using the fact that P, Q have matching first moment. Thus, as $F_P - F_Q$ integrates to 0, we can use (17) to see that

$$(I^1)^*(F_P - F_Q)(x) = \int_x^\infty (F_P - F_Q)(t)\, dt.$$

Assume the result holds for $k-1$. We claim that $(I^{k-1})^*(F_P - F_Q)$ integrates to 0, which is true as

$$\langle 1, (I^{k-1})^*(F_P - F_Q)\rangle = \langle 1, (I^k)^*(dP - dQ)\rangle = \langle x^k/k!, dP - dQ\rangle = 0,$$

the last step using the fact that P, Q have matching kth moment. Hence, as $(I^{k-1})^*(F_P - F_Q)$ integrates to 0, we can use (17) and conclude that

$$\begin{aligned} (I^k)^*(F_P - F_Q)(x) &= (I^1)^*(I^{k-1})^*(F_P - F_Q)(x) \\ &= \int_x^\infty (I^{k-1})^*(F_P - F_Q)(t)\, dt \\ &= \int_x^\infty \int_{t_k}^\infty \cdots \int_{t_2}^\infty (F_P - F_Q)(t_1)\, dt_1 \cdots dt_k, \end{aligned}$$

where in the last step we used the inductive hypothesis. This completes the proof.

References

Theodore W. Anderson and Donald A. Darling. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2):193–212, 1952.

Theodore W. Anderson and Donald A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

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.

K. Babenko. *Theoretical Foundations and Construction of Numerical Algorithms for the Problems of Mathematical Physics*. 1979. In Russian.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

Bhaswar B. Bhattacharya. *Power of graph-based two-sample tests*. PhD thesis, Stanford University, 2016.

M. Birman and M. Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Mathematics of the USSR-Sbornik*, 73(115):331–335, 1967. In Russian.

Maurice C. Bryson. Heavy-tailed distributions: Properties and tests. *Technometrics*, 16(1):61–68, 1974.

Fred Calitz. An alternative to the Kolmogorov-Smirnov test for goodness of fit. *Communications in Statistics: Theory and Methods*, 16(12):3519–3534, 1987.

Richard M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7(4):697–717, 1979.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schelkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Norbert Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Annals of Statistics*, 16(2):772–783, 1988.

Arnold Janssen. Global power functions of goodness of fit tests. *Annals of Statistics*, 28(1):239–253, 2000.

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

Regina Y. Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993.

Enno Mammen. Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics*, 19(2): 741–759, 1991.

David M. Mason and John H. Schuenemeyer. A modified Kolmogorov-Smirnov test sensitive to tail alternatives. *Annals of Statistics*, 11(3):933–946, 1983.

Alfred Muller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Yurii Nesterov. *Squared Functional Systems and Optimization Problems*, pages 405–440. Springer, 2000.

Aaditya Ramdas, Sashank Reddi, Barnabas Pczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *Twenty-Ninth Conference on Artificial Intelligence*, pages 3571–3577, 2015a.

Aaditya Ramdas, Sashank Reddi, Barnabas Poczos, Aarti Singh, and Larry Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*, 2015b.

Michael I. Rosen. Niels Hendrik Abel and equations of the fifth degree. *The American Mathematical Monthly*, 102(6):495–505, 1995.

Paul R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B*, 67(4):515–530, 2005.

Mark F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2291, 2013.

Naum Z. Shor. *Nondifferentiable Optimization and Polynomial Problems*. Nonconvex Optimization and Its Applications. Springer, 1998.

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

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Scholkopf, and Gert R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Gabor J. Szekely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10): 1249–1272, 2004.

Aad van der Vaart and Jon Wellner. *Weak Convergence*. Springer, 1996.

Yu-Xiang Wang, Alexander Smola, and Ryan J. Tibshirani. The falling factorial basis and its statistical applications. *International Conference on Machine Learning*, 31, 2014.