Supplement to "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

This supplementary document contains additional details, proofs, and experiments for the paper "A Higher-Order Kolmogorov-Smirnov Test". All section, figure, and equation numbers in this document begin with the letter "S", to differentiate them from those appearing in the main paper (which appear without the prepended letter "S").

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

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

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

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 $t \leq 0$.

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.

S.2 Proof of Proposition 1

We first claim that $F(x) = |x|^k/k!$ is an envelope function for $\mathcal{F}_k$, meaning $f \leq F$ 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 $|f^{(k)}(x)| \leq \mathrm{TV}(f^{(k)}) \leq 1$; 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 $\mathrm{TV}((f/R)^{(k)}) \leq 1$. 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

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

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 S.1 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 $P = Q$.

S.3 Statement and Proof of Lemma S.1

Lemma S.1. *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 $P(A) = Q(A)$. □

S.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

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

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 $Z^0_{(N)}$,

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

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

S.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

$$p(x) = \begin{cases} s(x) + (x-a)(b-x)t(x) & k \text{ even} \\ (x-a)s(x) + (b-x)t(x) & k \text{ odd} \end{cases}, \tag{S.2}$$

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).

S.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

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

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

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

Therefore

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

as desired.

S.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

$$\|g_{t_i} - g_{t_{i+1}}\|_2 \leq \epsilon, \quad i \in \{0\} \cup [R], \tag{S.3}$$

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

$$\begin{aligned} k!^2 \|g_s - g_t\|_2^2 &\leq k^2 (t-s)^2 \bigg(\int_s^\infty (x-s)^{2k+\delta} \, dP(x) \bigg)^{1/p} \bigg(\int_s^\infty 1^q \, (x) dP \bigg)^{1/q} \\ &\leq \frac{M^{1/p}}{(k-1)!^2} (t-s)^2, \end{aligned} \tag{S.4}$$

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

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

$$\begin{aligned} k!^2 \|g_t - 0\|_2^2 &= \int_t^\infty (x-t)^{2k} \, dP(x) \\ &\leq \bigg(\int_t^\infty (x-t)^{2k+\delta} \, dP(x) \bigg)^{2k/(2k+\delta)} \big(P(X \geq t)\big)^{\delta/(2k+\delta)} \\ &\leq M^{2k/(2k+\delta)} \bigg(\frac{\mathbb{E}|X|^{2k+\delta)}}{t^{2k+\delta}} \bigg)^{\delta/(2k+\delta)} = \frac{M}{t^\delta}, \end{aligned} \tag{S.5}$$

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 (S.4), to meet (S.3), 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 (S.5), 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

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

and (S.4), (S.5) imply that we have met (S.3). 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.

S.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 $1/\epsilon^2$).

S.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.

S.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

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

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

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

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.

S.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 $(\mathbb{P}_m - \mathbb{Q}_n)f = |(\mathbb{P}_m - \mathbb{Q}_n)g_{mn}|$. (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

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

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 $m, n \to \infty$,

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

because $t_{mn} \to t$ by definition, and $\mathbb{P}_m|x|^{k-1} \to \mathbb{P}|x|^k$. 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 $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} = (\mathbb{P}_m - \mathbb{Q}_n)f$ for each m, n, we must have $(\mathbb{P} - \mathbb{Q})g = (\mathbb{P} - \mathbb{Q})f$, 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 $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} = (\mathbb{P}_m - \mathbb{Q}_n)f$ 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 $\to \infty$, so again we have a representer in $\mathcal{G}_k$, as desired.

S.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.

S.13 Additional Experiments

S.13.1 Local Density Differences Continued

Figure S.1 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.

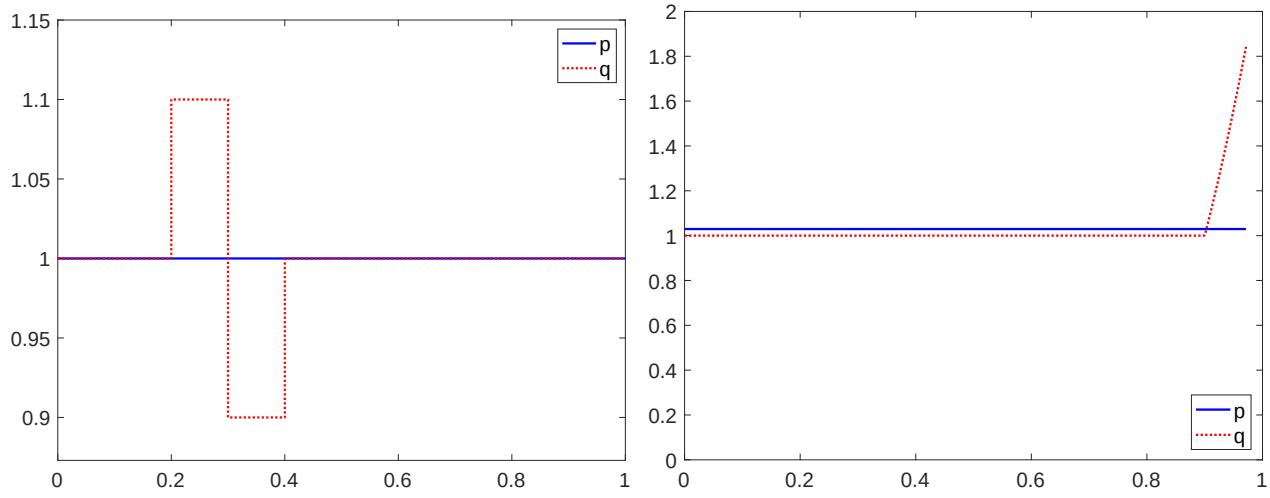

Figure S.1: *Densities for the local density difference experiments.*

S.13.2 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 S.2 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.

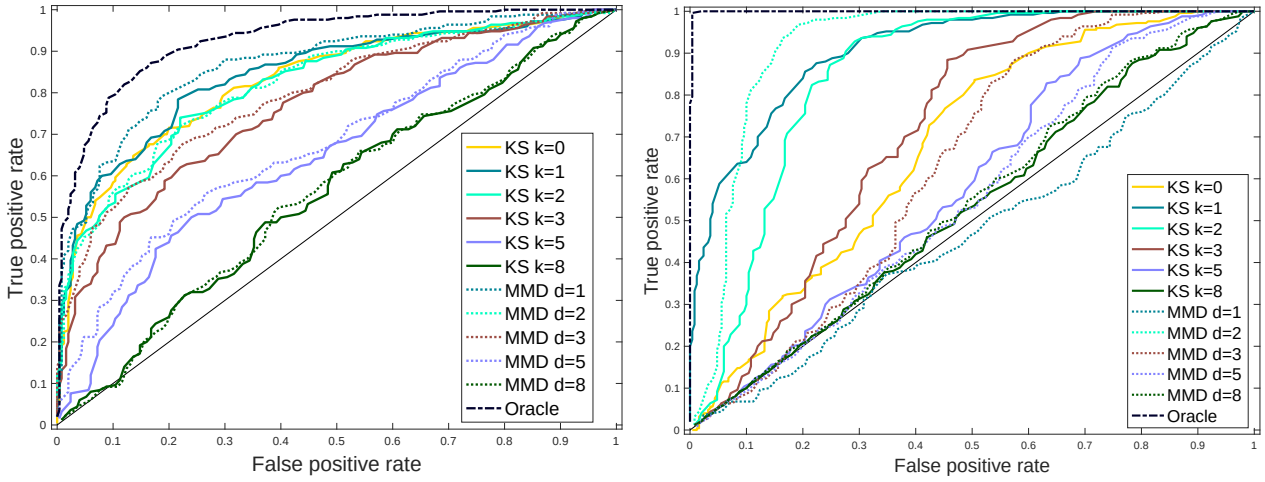

Figure S.2: *ROC curves for $P = N(0,1)$, $Q = N(0.2,1)$ (left), and $P = N(0,1)$, $Q = t(3)$ (right).*

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.

S.14 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

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

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

$$\begin{aligned}
\sup_{f \in \mathcal{F}_k} |\mathbb{P}f - \mathbb{Q}f| &= \sup_{f \in \mathcal{F}_k} |\langle f, dP - dQ \rangle| \\
&= \sup_{f \in \mathcal{F}_k} |\langle I^k D^k f, dP - dQ \rangle| \\
&= \sup_{\substack{h : \mathrm{TV}(h) \leq 1, \\ h(0+) = 0 \text{ or } h(0-) = 0}} |\langle I^k h, dP - dQ \rangle| \\
&= \sup_{\substack{h : \mathrm{TV}(h) \leq 1, \\ h(0+) = 0 \text{ or } h(0-) = 0}} |\langle h, (I^k)^*(dP - dQ) \rangle| \\
&= \|(I^1)^* (I^k)^* (dP - dQ)\|_\infty. \qquad \text{(S.6)}
\end{aligned}$$

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

$$((I^1)^* g)(t) = \int_t^\infty g(x) dx, \quad t \in \mathbb{R}. \qquad \text{(S.7)}$$

Returning to (S.6), 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 (S.7), 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 (S.7) 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 (S.7) 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

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

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

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

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