Maximum Mean Discrepancy Meets Neural Networks: The Radon-Kolmogorov-Smirnov Test

Seunghoon Paik[1], Michael Celentano[1], Alden Green[2], Ryan J. Tibshirani[1]

[1]University of California, Berkeley, [2]Stanford University

Abstract

Maximum mean discrepancy (MMD) refers to a general class of nonparametric two-sample tests that are based on maximizing the mean difference between samples from one distribution P versus another Q, over all choices of data transformations f living in some function space $\mathcal{F}$. Inspired by recent work that connects what are known as functions of *Radon bounded variation* (RBV) and neural networks (Parhi and Nowak, 2021, 2023), we study the MMD defined by taking $\mathcal{F}$ to be the unit ball in the RBV space of a given smoothness degree $k \geq 0$. This test, which we refer to as the *Radon-Kolmogorov-Smirnov* (RKS) test, can be viewed as a generalization of the well-known and classical Kolmogorov-Smirnov (KS) test to multiple dimensions and higher orders of smoothness. It is also intimately connected to neural networks: we prove that the witness in the RKS test—the function f achieving the maximum mean difference—is always a ridge spline of degree k, i.e., a single neuron in a neural network. We can thus leverage the power of modern neural network optimization toolkits to (approximately) maximize the criterion that underlies the RKS test. We prove that the RKS test has asymptotically full power at distinguishing any distinct pair $P \neq Q$ of distributions, derive its asymptotic null distribution, and carry out experiments to elucidate the strengths and weaknesses of the RKS test versus the more traditional kernel MMD test.

1 Introduction

In this paper, we consider the fundamental problem of nonparametric two-sample testing, where we observe independent samples $x_i \sim P$, $i = 1, \ldots, m$ and $y_i \sim Q$, $i = 1, \ldots, n$, all samples assumed to be in $\mathbb{R}^d$, and we use the data to test the hypothesis

$$H_0 : P = Q, \quad \text{versus} \quad H_1 : P \neq Q.$$

Though there are many forms of nonparametric two-sample tests, we will focus on a class of tests based on *maximum mean discrepancy* (MMD), which measures distance between two sets of samples by taking the maximum difference in sample averages over a function class $\mathcal{F}$:

$$\rho(P_m, Q_n; \mathcal{F}) = \sup_{f \in \mathcal{F}} |P_m(f) - Q_n(f)|. \tag{1}$$

Here $P_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ is the empirical distribution and $P_m(f) = \frac{1}{m}\sum_{i=1}^m f(x_i)$ the corresponding empirical expectation operator based on x_i, $i = 1, \ldots, m$; and likewise for Q_n and $Q_n(f)$.

The discrepancy in (1) is quite general and certain choices of $\mathcal{F}$ recover well-known distances that give rise to two-sample tests. A noteworthy univariate ($d = 1$) example is the *Kolmogorov-Smirnov* (KS) distance (Kolmogorov, 1933; Smirnov, 1948). Though it is more typically defined in terms of cumulative distribution functions (CDFs) or ranks, the KS distance can be seen as the MMD that results when $\mathcal{F}$ is taken to be the class of functions with at most unit *total variation* (TV).

Some attempts have been made to define a multivariate KS distance based on multivariate CDFs (Bickel, 1969) or ranks (Friedman and Rafsky, 1979), but the resulting tests have a few limitations, such as being expensive to compute or exhibiting poor power in practice. In fact, even the univariate KS test is known to be insensitive to certain kinds of alternatives, such as those $P \neq Q$ which differ "in the tails" (Bryson, 1974).

This led Wang et al. (2014) to recently propose a higher-order generalization of the KS test, which uses the MMD based on a class of functions whose *derivatives* are of bounded TV.

In this work we propose and study a new class of MMDs based on a kind of multivariate total variation known as *Radon total variation* (RTV). The definition of RTV is necessarily technical, and we delay it until Section 2. For now, we remark that if $d = 1$ then RTV reduces to the usual notion of total variation, which means that our RTV-based MMD—the central focus of this paper—directly generalizes the KS distance to multiple dimensions. We refer to this RTV-based MMD as the *Radon-Kolmogorov-Smirnov* (RKS) distance, and the resulting two-sample test as the RKS test. We also will consider the MMD defined using a class of functions whose derivatives are of bounded RTV, which in turn generalizes the higher-order KS distance of Wang et al. (2014) to multiple dimensions. Sadhanala et al. (2019) recently showed that the higher-order KS test can be more sensitive to departures in the tails, and we will give empirical evidence that our higher-order RKS test can be similarly sensitive to tail behavior.

Our test statistic also has a simple but revealing characterization in terms of *ridge splines*. A ridge spline of integer degree $k \geq 0$ is defined for a direction $w \in \mathbb{S}^{d-1}$ and an offset $b \in \mathbb{R}$ as the truncated polynomial $f(x) = (w^\top x - b)_+^k$. Here we use $\mathbb{S}^{d-1}$ to denote the unit ℓ_2 sphere in in $\mathbb{R}^d$ (the set of vectors in with unit ℓ_2 norm), and we use the abbreviation $t_+ = \max\{t, 0\}$. We prove (Theorem 3) that a k^{th} degree ridge spline *witnesses* the k^{th} degree RKS distance: there is a k^{th} degree ridge spline f with unit Radon total variation for which $P_m(f) - Q_n(f)$ equals the RKS distance. Thus the RKS test statistic can be equivalently written as

$$T_{d,k} = \max_{(w,b)\in\mathbb{S}^{d-1}\times[0,\infty)} \left| \frac{1}{m}\sum_{i=1}^{m}(w^\top x_i - b)_+^k - \frac{1}{n}\sum_{i=1}^{n}(w^\top y_i - b)_+^k \right|. \tag{2}$$

Ridge splines—and in particular the first-degree ridge spline $f(x) = (w^\top x - b)_+$, which is also called a rectified linear (ReLU) unit—are fundamental building blocks of modern neural networks. From this perspective, the representation (2) says that the RKS distance is achieved by a single neuron in a two-layer neural network. In fact, the connections between Radon total variation and neural networks run deeper. The RKS distance can be viewed as finding the maximum discrepancy between sample means over all two-layer neural networks (of arbitrarily large width) subject to a weight decay constraint (Section 2). Although we focus throughout on two-layer neural networks, the MMD defined over deeper networks can be viewed the RKS distance in the representation learned by the hidden layers.

The representation (2) also hints at several interesting statistical properties. Let $X \sim P_m$ and $Y \sim Q_n$ be random variables drawn from the empirical distributions of $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^n$, respectively. The 0^{th} degree RKS distance finds the direction $w \in \mathbb{S}^{d-1}$ along which the usual Kolmogorov-Smirnov distance—maximum gap in CDFs—between the univariate random variables $w^\top X$ and $w^\top Y$ is as large as possible. For $k \geq 1$, the k^{th} degree statistic finds the direction $w \in \mathbb{S}^{d-1}$ along which the higher-order KS distance—maximum gap in truncated k^{th} moments—between $w^\top X$ and $w^\top Y$ is maximized.

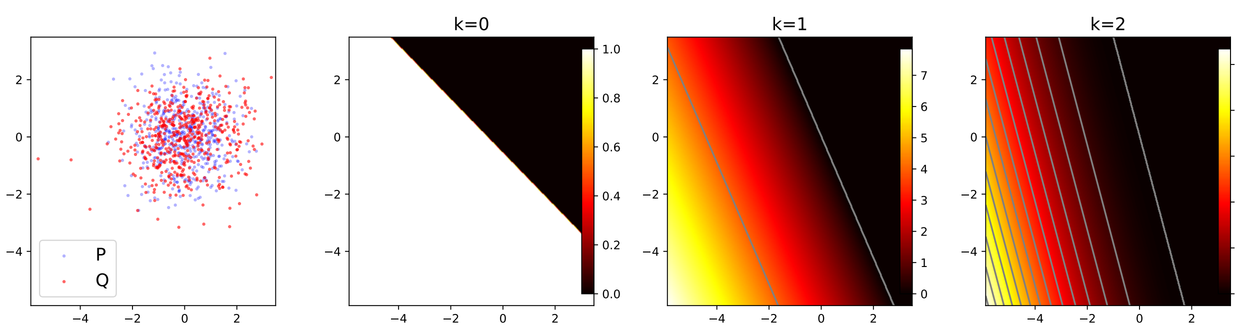


Figure 1: Illustration of RKS tests for $P = \mathcal{N}_2(0, I)$ and $Q = \mathcal{N}_2(0, D)$, where $D = \text{diag}(1.4, 1)$.

Thus, speaking informally, we expect the RKS test to be particularly sensitive to the case when the laws of $w^\top X$ and $w^\top Y$ differ primarily in just a few directions w (because we are taking a maximum over such directions in (2)). In other words, we expect the RKS test to be sensitive to *anisotropic* differences between P and Q. On the other hand, taking a larger value of k results in higher-order truncated sample moments, which can be much more sensitive to small differences in the tails. Figure 1 gives an illustrative example. We consider two normal distributions with equal means and covariances differing in just one direction (along the x-axis). The figure displays the witnesses (ridge splines) for the RKS test when $k = 0, 1, 2$. For larger k, the witness is more aligned with the direction in which P and Q vary, and also places more weight on the tails.

Summary of contributions. Our main contributions are as follows.

- We propose a new class of two-sample tests using a k^{th} degree Radon total variation MMD. We prove a representer theorem (Theorem 3) which shows that the MMD is witnessed by a ridge spline (establishing the representation in (2)). This connects our proposal to two-sample testing based on neural networks, and helps with optimization.
- Under the null $P = Q$, we derive the asymptotic distribution of the test statistic (Theorem 5).
- Under the alternative $P \neq Q$, we give upper bounds on the rate at which the test statistic concentrates around the population-level MMD. Using the fact that the population-level MMD is a metric, we then show that the RKS test is *consistent* against any fixed $P \neq Q$: it has asymptotic error (sum of type I and type II errors) tending to zero (Theorem 6).
- We complement our theory with numerical experiments to explore the operating characteristics of the RKS test compared to other popular nonparametric two-sample tests.

Related work. Recently, there has been a lot of interest in multivariate nonparametric two-sample tests using kernel MMDs (Gretton et al., 2012), and energy distances (Baringhaus and Franz, 2004; Székely and Rizzo, 2005). The latter are in many cases equivalent to a kernel MMD (Sejdinovic et al., 2013). It is worth being clear that the Radon bounded variation space is *not* an RKHS, and as such, our test cannot be seen as a special case of a kernel MMD. There are also many other kinds of multivariate nonparametric two-sample tests, such as graph-based tests using kNN graphs (Schilling, 1986; Henze, 1988) or spanning-trees (Friedman and Rafsky, 1979). The RKS test statistic (2) avoids the need to construct a graph over the samples, though we note that one could also use a graph to measure (approximate) multivariate total variation of a different variety (sometimes called *measure-theoretic* TV, which is not the same as Radon TV), and one could then form a related two-sample test, accordingly.

Some literature refers to maximum mean discrepancy (MMD) with respect to an arbitrary function class as an *integral probability metric* (IPM) (Müller, 1997). While IPMs look at differences in $\mathrm{d}P - \mathrm{d}Q$, closely related are ϕ-divergences, which look at differences in $\mathrm{d}P/\mathrm{d}Q$ (Sriperumbudur et al., 2009).

In machine learning, two-sample tests play a fundamental role in generative adversarial networks (GANs) (Goodfellow et al., 2014), which have proven to be an effective way to generate new (synthetic) draws that adhere to the distribution of a given high-dimensional set of samples. In short, GANs are trained to produce synthetic data which a two-sample test cannot distinguish from "real" samples. However, optimizing a GAN is notoriously unstable. It has been suggested that the stability of GANs can be improved if the underlying two-sample test is based on an MMD, with constraints on, e.g., the Lipschitz constant (Arjovsky et al., 2017), RKHS norm (Li et al., 2017), or Sobolev norm (Mroueh et al., 2018) of the witness. Unlike Radon TV, these function classes do not have an intrinsic connection to neural networks, which in practice are most commonly used to train the discriminator in a GAN.

At a technical level, our paper falls into a line of work exploring neural networks from the functional analytic perspective. The equivalences between weight decay of parameters in a two-layer neural network and first-degree ($k = 1$) Radon total variation were discovered in Savarese et al. (2019); Ongie et al. (2020), and further formalized and extended by Parhi and Nowak (2021) to the higher-degree case ($k > 1$). We make use of these equivalences to prove results on the Radon TV MMD. Lastly, we note that the sensitivity of neural networks to variation in only a few directions was observed even earlier, in Bach (2017).

2 The Radon-Kolmogorov-Smirnov test

For independent samples $x_i \sim P$, $i = 1, \ldots, m$ and $y_i \sim Q$, $i = 1, \ldots, n$, and a given integer $k \geq 0$, the RKS test statistic $T_{d,k}$ is defined as in (2). This statistic searches for a marginal (defined by projection onto the direction w) with the largest discrepancy in a truncated k^{th} moment. As with any two-sample test, we can calibrate the rejection threshold for the RKS test statistic using a permutation null. Given any user-chosen level $\alpha \in [0,1]$, we reject $H_0 : P = Q$ if

$$\frac{\sum_{j=1}^{B} \mathbb{1}\{T_{d,k}(z_{\pi_j}) \geq T_{d,k}(z)\} + 1}{B+1} \leq \alpha. \tag{3}$$

Here we use $z = (x_1, \dots, x_m, y_1, \dots, y_n) \in \mathbb{R}^{m+n}$ to denote the original data sequence, and we use $T_{d,k}(z)$ to emphasize that the statistic in (2) is computed on z. Further, each π_j is a permutation of $\{1, \dots, n\}$, drawn uniformly at random, and we use $T_{d,k}(z_{\pi_j})$ to denote the test statistic computed on the permuted sequence $z_{\pi_j} = (z_{\pi(1)}, \dots, z_{\pi(m+n)})$. Standard results on permutation tests imply that the test defined by the rejection threshold (3) has type I error control at the level α.

Finding the value of w and b achieving the supremum in (2) is a nonconvex problem. However, it bears a connection to neural network optimization, which is of course among the most familiar and widely-studied nonconvex problems in machine learning. We show in Corollary 1 that problem (2) can be equivalently cast as finding the maximum discrepancy between sample means over all two-layer neural networks (of arbitrarily large width) under a weight decay-like constraint. In particular, if we denote by $f_{(a_j,w_j,b_j)_{j=1}^N}$ the N-neuron two-layer neural network defined by

$$f_{(a_j,w_j,b_j)}(x) = a_j(w_j^\mathsf{T} x - b_j)_+^k \quad \text{and} \quad f_{(a_j,w_j,b_j)_{j=1}^N}(x) = \sum_{j=1}^{N} f_{(a_j,w_j,b_j)}(x), \tag{4}$$

then the RKS test statistic is equivalently

$$\begin{aligned} T_{d,k} = \max_{f = f_{(a_j,w_j,b_j)_{j=1}^N}} \quad & P_m(f) - Q_n(f) \\ \text{subject to} \quad & \textstyle\sum_{j=1}^{N} |a_j| \|w_j\|_2^k \leq 1, \\ & b_j \geq 0, \ j = 1, \dots, N. \end{aligned} \tag{5}$$

This equivalence holds for any number of neurons $N \geq 1$, which is why we say that the network can be of "arbitrarily large width".

In Section 4, we use a Lagrangian form of this constrained optimization, allowing us to leverage existing deep learning toolkits to compute the RKS distance. As with neural network optimization in general, we cannot prove that any practical first-order scheme—such as gradient descent or its variants—is able to find the global optimum of (5), or its Lagrange form. However, we find empirically that the test statistic resulting from such first-order schemes has favorable behavior in practice, such as stability (with respect to the choice of learning rate) and strong power (with respect to certain classes of alternatives). Finally, we note that in practice we use the permutation threshold (3) to calibrate the statistic computed by first-order optimization, i.e., the same first-order optimization scheme is applied to the original data sequence and to every permuted sequence. In effect, we can view this as *redefining* the RKS test to be the output of such a first-order scheme, and as the permutation null controls type I error for any measurable function of data, it also controls type I error for a test statistic computed this way.

2.1 Functions of Radon bounded variation

The RKS distance, like the univariate KS distance, can be viewed as the maximum mean discrepancy with respect to a function class defined by a certain notion of smoothness. Whereas the univariate KS distance is the MMD over functions with total variation is bounded by 1, the RKS distance is the MMD over functions with Radon total variation bounded by 1, subject to a boundary condition. This will be shown a bit later; first, we must cover preliminaries needed to understand Radon total variation.

Heuristically, the k^{th} degree Radon total variation of $f : \mathbb{R}^d \to \mathbb{R}$ measures the average smoothness of the "marginals" of f. To be more concrete, consider a function $f : \mathbb{R}^d \to \mathbb{R}$ which is infinitely differentiable, and whose derivatives of any order decay super-polynomially: $f \in C^\infty(\mathbb{R}^d)$ and $\sup_{x \in \mathbb{R}^d} |x^\alpha \partial^\beta f(x)| < \infty$. Here

$\alpha = (\alpha_1, \ldots, \alpha_d)$ and $\beta = (\beta_1, \ldots, \beta_d)$ are multi-indices (with nonnegative integer elements), and we use the abbreviations

$$x^\alpha = \prod_{i=1}^d x_i^{\alpha_i}, \quad \text{and} \quad \partial^\beta f = \partial_{x_1}^{\beta_1} \cdots \partial_{x_d}^{\beta_d} f.$$

The set of such functions is called the *Schwartz class*, and denoted by $\mathcal{S}(\mathbb{R}^d)$. The *Radon transform* of a function $f \in \mathcal{S}(\mathbb{R}^d)$ is itself another function $\mathcal{R}\{f\} : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ defined as

$$\mathcal{R}\{f\}(w, b) = \int_{w^\mathsf{T} x = b} f(x)\, \mathrm{d}x,$$

where the integral is with respect to the Lebesgue surface measure on the hyperplane $\{w^\mathsf{T} x = b\}$. As a function of b, the Radon transform can be viewed as the marginal of the function f in the direction w. The k^{th} degree Radon total variation of f is then defined to be (Parhi and Nowak, 2021, 2023):[1]

$$\|f\|_{\mathrm{RTV}^k} = c_d \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \left| \partial_b^{k+1} \Lambda^{d-1} \mathcal{R}\{f\}(w, b) \right| \mathrm{d}\sigma(w, b),$$

where σ is the Hausdorff measure on $\mathbb{S}^{d-1} \times \mathbb{R}$, $c_d = 1/(2(2\pi)^{d-1})$, and Λ^{d-1} is the "ramp filter" defined by $\Lambda^{d-1} = (-\partial_b^2)^{\frac{d-1}{2}}$ with fractional derivatives interpreted in terms of Riesz potentials (see Parhi and Nowak (2021, 2023) for details). Equivalently, if we define an operator by $R_k\{f\}(w, b) = c_d \partial_b^{k+1} (-\partial_b^2)^{\frac{d-1}{2}} \mathcal{R}\{f\}(w, b)$, then we can write k^{th} degree Radon TV of f simply as $\|f\|_{\mathrm{RTV}^k} = \|R_k\{f\}\|_{L^1}$. The mapping $f \mapsto \|f\|_{\mathrm{RTV}^k}$ is a seminorm, which we call the RTV^k seminorm.

Parhi and Nowak (2021, 2023) and Parhi (2022) extend the RTV^k seminorm beyond Schwartz functions using functional analytic tools based on duality. The space of functions with bounded RTV^k seminorm is referred to as the RBV^k space (where RBV stands for "Radon bounded variation"). After this extension, one can still think of RTV^k seminorm as measuring average higher-order smoothness of the marginals of f, but it does not require that all derivatives and integrals exist in a classical sense. We refer the reader especially to Parhi (2022) for the complete functional analytic definition of the RBV^k space. For our purposes, it is only important that RBV^k functions enjoy the representation established by the following theorem. Note that in order to define the RBV^k space, Parhi and Nowak (2021); Parhi (2022) must view it as a space of equivalence classes of functions under the equivalence relation $f \sim g$ if $f(x) = g(x)$ for Lebesgue almost every $x \in \mathbb{R}^d$. We denote elements of RBV^k as f, and the functions they contain as $f \in \mathsf{f}$.

Proposition 1 (Adaptation of Theorem 22 in Parhi and Nowak (2021); Theorem 3.8 in Parhi (2022))**.** *Fix any $k \geq 0$. For each $\mathsf{f} \in \mathrm{RBV}^k$, there exists a representative $f \in \mathsf{f}$ which satisfies for all $x \in \mathbb{R}^d$*

$$f(x) = \int_{\mathbb{S}^{d-1} \times [0, \infty)} (w^\mathsf{T} x - b)_+^k \, \mathrm{d}\mu(w, b) + q(x),$$

for some finite signed Borel measure μ on $\mathbb{S}^{d-1} \times [0, \infty)$ satisfying $\|\mu\|_{\mathrm{TV}} = \|\mathsf{f}\|_{\mathrm{RTV}^k}$, and some polynomial q of degree at most k, where we use $\|\cdot\|_{\mathrm{TV}}$ for the total variation norm in the sense of measures.

Proposition 1 is the key result connecting the RBV^k space to two-layer neural networks. It states that any RBV^k functions is equal almost everywhere to the sum of a (possibly infinite-width) two-layer neural network and a polynomial. Furthermore, the ℓ_1 norm of coefficients in the last layer of this neural network (formally, the total variation norm $\|\mu\|_{\mathrm{TV}}$ of the measure μ defined over the coefficients) equals the RBV^k seminorm of the function in question. As stated, Proposition 1 is a slight refinement of results in Parhi and Nowak (2021); Parhi (2022), which we show in Appendix A.

Parhi and Nowak (2021) consider the nonparametric regression problem defined by minimizing, over all functions f, the squared loss incurred by f with respect to a given finite data set plus a penalty on $\|f\|_{\mathrm{RTV}^k}$. They show this is solved by the sum of a *finite-width* two-layer neural network and a polynomial, and hence provide a representation theorem for RTV^k-penalized nonparametric regression analogous to that for kernel

[1] A remark on notation and nomenclature: we use RBV^k to refer to the space which Parhi and Nowak (2021); Parhi (2022); Parhi and Nowak (2023) instead call RBV^{k+1}. That is, we index based on the *degree* of the underlying ridge spline, whereas the aforementioned papers index based on the *order*, traditionally defined as the degree plus one, $m = k + 1$.

ridge regression (Wahba, 1990). In a coming subsection, we will establish a similar connection in the context of two-sample testing: the MMD under an RTV^k constraint is realized by a single neuron.

Before moving on, it is worth providing some intuition for the results just discussed. First, the operator R_k should be understood as transforming a function f into the Radon domain: namely, it specifies f via the higher-order derivatives of its marginals. Then, the RTV^k acts as the L^1 norm in the Radon domain. Lastly, the reason two-layer neural networks and ridge splines emerge as the solutions to nonparametric regression and maximum mean discrepancy problems under an RTV^k constraint is that ridge splines are *sparse* in the Radon domain. In particular, for $f(x) = (w^\mathsf{T} x - b)_+^k$, we have (Parhi and Nowak, 2021):

$$R_k\{f\}(w, b) = \frac{1}{2}\big(\delta_{(w,b)} + (-1)^{k+1}\delta_{(-w,-b)}\big),$$

where $\delta_{(\pm w, \pm b)}$ denote point masses at $(\pm w, \pm b)$. Thus, RTV^k-penalized optimization problems are solved by two-layer neural networks due to the tendency of the L^1 norm to encourage sparse solutions.

2.2 Pointwise evaluation of RBV functions

Recall, the RBV^k space contains equivalence classes of functions whose members differ on sets of Lebesgue measure zero. Therefore, at face value, point evaluation is meaningless for elements of the RBV^k space. This poses a problem for defining an MMD with respect to RBV functions, since the MMD depends on empirical averages, which require function evaluations over samples. Fortunately, we show that there is a natural way to resolve point evaluation over the RBV^k space: each equivalence class in RBV^k has, possibly subject to a mild boundary condition, a unique representative which satisfies a suitable notion of continuity. We can then define point evaluation with respect to this representative.

For the case $k = 0$, we need to introduce a new notion of continuity, which we call *radial cone continuity*. For $k \geq 1$, we rely on the standard notion of continuity. Radial cone continuity is defined as follows.

Definition 1. *A function f is said to be* radially cone continuous *if it is continuous at the origin, and for any $x \in \mathbb{R}^d \setminus \{0\}$, we have $f(x + \epsilon v) \to f(x)$ whenever (i) $\epsilon \to 0$ and (ii) $v \in \mathbb{S}^{d-1}$ and $v \to x/\|x\|_2$.*

Radial cone continuity can be viewed as a generalization of right-continuity to higher dimensions. Indeed a consequence of radial cone continuity is that for every $v \in \mathbb{R}^d$, the restriction of the function f to the ray emanating from the origin in the direction v is right-continuous (i.e., $t \mapsto f(tv)$, $t > 0$ is right-continuous). The concept of radial cone continuity, however, is stronger than right-continuity along rays—it additionally requires continuity along any path which is tangent to and approaches x from the same direction as the ray that points from the origin to x. The following theorem shows that (subject to a boundary condition for the case $k = 0$) every equivalence class $\mathsf{f} \in \mathrm{RBV}^k$ contains a unique continuous representative when $k \geq 1$, and radially cone continuous representative when $k = 0$.

Theorem 1. *For $k = 0$, if $\mathsf{f} \in \mathrm{RBV}^0$ contains a function that is continuous at the origin, then it contains a unique representative that is radially cone continuous. Moreover, for $k \geq 1$, each $\mathsf{f} \in \mathrm{RBV}^k$ contains a unique representative that is continuous and this representative is in fact $(k-1)$-times continuously differentiable.*

We prove Theorem 1 in Appendix B. The theorem motivates the definition of the function spaces:

$$\begin{aligned}
\mathrm{RBV}_c^0 &= \Big\{ f \,:\, f \in \mathsf{f} \text{ for some } \mathsf{f} \in \mathrm{RBV}^0,\ f \text{ radially cone continuous} \Big\}, \\
\mathrm{RBV}_c^k &= \Big\{ f \,:\, f \in \mathsf{f} \text{ for some } \mathsf{f} \in \mathrm{RBV}^k,\ f \text{ continuous} \Big\}, \quad k \geq 1.
\end{aligned} \tag{6}$$

Unlike RBV^k, the space RBV_c^k contains functions rather than equivalence classes of functions, so that point evaluation is well-defined. The Radon total variation of a member $f \in \mathsf{f}$ is defined in the natural way as the Radon total variation of the function class in RBV^k to which it belongs.

2.3 The RKS distance is an MMD

We are ready to show that the RKS distance is equivalent to the MMD over functions f with $\|f\|_{\mathrm{RTV}^k} \leq 1$, subject to a boundary condition. In this sense, it can be viewed as a generalization of the classical univariate

KS distance, which, as we have mentioned, is the MMD over the space of functions with total variation at most 1. Precisely, the RKS distance is the MMD over the function space

$$\mathcal{F}_k = \Big\{ f \in \mathrm{RBV}_c^k \,:\, \|f\|_{\mathrm{RTV}^k} \le 1,\ \partial^\alpha f(0) \text{ exists and equals } 0 \text{ for all } |\alpha| \le k \Big\}. \tag{7}$$

The boundary condition $\partial^\alpha f(0) = 0$ for all multi-indices $\alpha = (\alpha_1, \dots, \alpha_d)$ with $|\alpha| = \sum_{j=1}^d \alpha_j \le k$ is needed to guarantee that the MMD is finite. We can interpret this condition intuitively as requiring the polynomial q to be zero in the representation in Proposition 1. Otherwise, q would be able to grow without bound under the constraint $\|f\|_{\mathrm{RTV}^k} \le 1$, which we could then use to drive the MMD criterion to ∞, provided that P and Q had any moment differences (in their first k moments).

The following representation theorem proves that the RKS distance is the MMD over (7).

Theorem 2. *Fix any $k \ge 0$. For any $f \in \mathcal{F}_k$, there exists a finite, signed Borel measure μ on $\mathbb{S}^{d-1} \times [0, \infty)$ such that for all $x \in \mathbb{R}^d$,*

$$f(x) = \int_{\mathbb{S}^{d-1} \times [0,\infty)} (w^\mathsf{T} x - b)_+^k \, \mathrm{d}\mu(w, b),$$

and $\|\mu\|_{\mathrm{TV}} = \|f\|_{\mathrm{RTV}^k}$. When $k = 0$, the measure μ may be taken to be supported on $\mathbb{S}^{d-1} \times (0, \infty)$.

We prove Theorem 2 in Appendix C. Note that the representation provided by Theorem 2 is like that in Proposition 1 except that the integral is restricted to $b \ge 0$, and the polynomial term is $q = 0$. The latter is a consequence of the boundary condition imposed in the definition of $\mathcal{F}_k$.

The next theorem establishes the equivalence between $T_{d,k}$ as defined in (2) and the MMD over $\mathcal{F}_k$. Its proof is in Appendix D.

Theorem 3. *Fix any $k \ge 0$. Define*

$$\mathcal{G}_k = \Big\{ (w^\mathsf{T} \cdot - b)_+^k \,:\, (w, b) \in \mathbb{S}^{d-1} \times [0, \infty) \Big\},$$

where by convention we take $t_+^0 = 1\{t \ge 0\}$. Then for any P and Q with finite k^{th} moments, we have

$$\rho(P, Q; \mathcal{F}_k) = \rho(P, Q; \mathcal{G}_k).$$

In particular, this implies that for the empirical distributions P_m and Q_n over any sets of samples $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^n$, we have $\rho(P_m, Q_n; \mathcal{F}_k) = T_{d,k}$, where the latter is as defined in (2).

Lastly, as a consequence of the above theorem, we establish the representation claimed earlier in (5). Its proof is simple enough that we present it here.

Corollary 1. *Fix any $k \ge 0$. For the empirical distributions P_m and Q_n over any sets of samples $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^n$, the RKS test statistic $T_{d,k}$ defined in (2) is equivalent to that in (4), (5), for any $N \ge 1$.*

Proof. For $\|w\|_2 = 1$ and $b \in \mathbb{R}$, Parhi and Nowak (2021) show that $\|(w^\mathsf{T} \cdot - b)_+^k\|_{\mathrm{RTV}^k} = 1$. Then, for any $w \in \mathbb{R}^d$ and $a \in \mathbb{R}$, by the k^{th} order homogeneity of the ridge splines, $\|a(w^\mathsf{T} \cdot - b)_+^k\|_{\mathrm{RTV}^k} = |a| \|w\|_2^k$. By the triangle inequality,

$$\|f_{(a_j, w_j, b_j)_{j=1}^N}\|_{\mathrm{RTV}^k} \le \sum_{j=1}^N |a_j| \|w_j\|_2^k.$$

Thus, defining

$$\mathcal{F}_{k,N} = \Big\{ f_{(a_j, w_j, b_j)_{j=1}^N} \,:\, (a_j, w_j, b_j) \in \mathbb{R} \times \mathbb{R}^d \times (0, \infty),\ \sum_{j=1}^N |a_j| \|w_j\|_2^k \le 1 \Big\},$$

we have $\mathcal{F}_{k,N} \subseteq \mathcal{F}_k$. Defining also $\mathcal{G}_k^+ = \{(w^\mathsf{T} \cdot - b)_+^k \,:\, (w, b) \in \mathbb{S}^{d-1} \times (0, \infty)\}$, we have $\mathcal{G}_k^+ \subseteq \mathcal{F}_{k,N}$, so

$$\sup\nolimits_{f \in \mathcal{G}_k^+} |P_m(f) - Q_n(f)| \le \sup\nolimits_{f \in \mathcal{F}_{k,N}} |P_m(f) - Q_n(f)| \le \sup\nolimits_{f \in \mathcal{F}_N} |P_m(f) - Q_n(f)|.$$

The leftmost supremum is unchanged if we include the origin, i.e., take the suprermum over $f \in \mathcal{G}_k$, which precisely defines $T_{d,k}$. By Theorem 3, this is equal to $\sup_{f \in \mathcal{F}_k} |P_m(f) - Q_n(f)|$, the rightmost supremum above, so we conclude that all inequalities are equalities, and $T_{d,k} = \sup_{f \in \mathcal{F}_{k,N}} |P_m(f) - Q_n(f)|$. Finally, we can remove the absolute value sign in the criterion because $\mathcal{F}_{k,N}$ is symmetric ($f \in \mathcal{F}_{k,N} \iff -f \in \mathcal{F}_{k,N}$), which completes the proof of (5) as an equivalent representation. ∎

2.4 The RKS distance identifies the null

An important property of the RKS distance for the purposes of two-sample testing is its ability to distinguish any two distinct distributions. In particular, we have the following result, whose proof is in Appendix E.

Theorem 4. *For any P, Q with finite k^{th} moments, $\rho(P, Q; \mathcal{F}_k) = 0$ if and only if $P = Q$, and is positive otherwise.*

Theorem 4 states that the RKS distance, in the population, perfectly distinguishes the null $H_0 : P = Q$ from the alternative $H_1 : P \neq Q$. Thus, the function space $\mathcal{F}_k$ is rich enough to make the MMD a *metric* at the population level. As an interesting contrast, we remark that this is not true of the kernel MMD with a polynomial kernel of degree k, as studied in Gretton et al. (2012). Indeed, if P and Q have the same moments up to order k (where "moments" is to be interpreted in the appropriate multivariate sense) then the kernel MMD in the population between P and Q is 0. In other words, we can see that the use of *truncated* moments, as given by averages of ridge splines, is the key to the discrepancy being a metric.

3 Asymptotics

Given any probability distribution P on $\mathbb{R}^d$ with finite $(2k+\Delta)$-moments for some $\Delta > 0$, let us define the corresponding Gaussian process $\mathbb{G}_P = \{G_{w,b} : (w, b) \in \mathbb{S}^{d-1} \times [0, \infty)\}$ by

$$G_{w,b} \sim \mathcal{N}\Big(0,\, \mathbb{E}_{x\sim P}\big[(w^\mathsf{T} x - b)_+^{2k}\big]\Big), \quad \mathrm{Cov}(G_{w,b}, G_{w',b'}) = \mathbb{E}_{x\sim P}\Big[(w^\mathsf{T} x - b)_+^k (w'^\mathsf{T} x - b')_+^k\Big]. \tag{8}$$

The importance of this Gaussian process is explained in the next result, which shows that its supremum provides the asymptotic null distribution of the RKS test statistic.

Theorem 5. *Fix any $k \geq 0$. Assume that P has finite $(2k + \Delta)$-moments for some $\Delta > 0$. If $k = 0$, then additionally assume that P has bounded density with respect to Lebesgue measure. When $P = Q$, we have*

$$\sqrt{\frac{mn}{m+n}}\, T_{d,k} \xrightarrow{d} \sup_{(w,b)\in\mathbb{S}^{d-1}\times[0,\infty)} |G_{w,b}|, \quad \text{as } m, n \to \infty.$$

Moreover, with a proper rejection threshold, the RKS test can have asymptotically zero type I error and full power against any fixed $P \neq Q$.

Theorem 6. *Fix any $k \geq 0$. Assume P, Q have finite $(2k + \Delta)$-moments for some $\Delta > 0$. If $k = 0$, then additionally assume that P, Q have bounded densities with respect to Lebesgue measure. Let $p = m + n$ and consider any sequence t_p such that $t_p \to 0$ and $t_p\sqrt{p} \to \infty$ as $m, n \to \infty$. Then the test which rejects when $T_{d,k} > t_p$ rejects with asymptotic probability 0 if $P = Q$ and asymptotic probability 1 if $P \neq Q$.*

We prove Theorems 5 and 6 in Appendices F and G. Roughly, Theorem 5 follows from general uniform central limit theorem results in Dudley (2014), after controlling the bracketing number of the function class $\mathcal{G}_k$. Theorem 6 follows from tail bounds that we establish for the concentration of the RKS distance around its population value, combined with the fact that the population distance is a metric (Theorem 4).

4 Experiments

We conduct experiments to compare the power of the RKS test with other multivariate nonparametric tests in multiple settings (different pairs of P and Q), both to investigate the strengths and weaknesses of the RKS test, and to investigate its behavior as we vary the smoothness degree k.

4.1 Computation of the RKS distance

First, we discuss computation of the RKS distance, as initially defined in (2). This a difficult optimization problem because of the nonconvexity of the criterion and the unit ℓ_2 norm constraint on w. Though we will not be able to circumvent the challenge posed by nonconvexity entirely, we can ameliorate it by moving to the overparametrized formulation in (5): this now casts the same computation as an optimization over the N-neuron neural network f in (4), subject to a constraint on $\sum_{j=1}^N |a_j| \|w_j\|_2^k$ (which is sometimes called the "path norm" of f). We call this problem "overparametrized" since any individual pair (w_j, b_j) would be sufficient to obtain the global optimum in (2), and yet we allow the optimization to search over N such pairs (w_j, b_j), $j = 1, \ldots, N$. Perhaps unsurprisingly, we find that overparametrization (larger N) generally makes optimization easier—more stable with respect to the initialization and choice of learning rate, discussed in more detail in Appendix J.

Problem (5) is still more difficult than we would like it to be, i.e., not within scope for typical first-order optimization techniques, due to the path norm constraint $\sum_{j=1}^N |a_j| \|w_j\|_2^k \leq 1$. Fortunately, it is easy to see that any bound here suffices to compute the RKS distance: we can instead use $\sum_{j=1}^N |a_j| \|w_j\|_2^k \leq C$, for any $C > 0$, rescale the criterion $P_m(f) - Q_n(f)$ by $1/C$, and then the value of the criterion at its optimum (the MMD value) is unchanged. This motivates us to instead solve a Lagrangian version of (5):

$$\min_{\substack{f = f_{(a_j, w_j, b_j)_{j=1}^N} \\ b_j \geq 0,\, j=1,\ldots,N}} \; -\frac{1}{k} \log \left(\frac{1}{N} \big| P_m(f) - Q_n(f) \big| \right) + \frac{\lambda}{kN} \sum_{i=1}^N |a_i| \|w_i\|_2^k, \tag{9}$$

which is the focus in all of our experiments henceforth. The use of a log transform for the mean difference in (9) is used because we find that it leads to a greater stability with respect to the choice of learning rate, as discussed in Appendix I. Importantly, the choice of Lagrangian parameter $\lambda > 0$ in (9) can be arbitrary, by the same rescaling argument as explained above: given any (local) optimizer f^* in (9), we can simply return the mean difference over the rescaled witness $f^* / \|f^*\|_{\mathrm{RTV}^k}$, where $\|f^*\|_{\mathrm{RTV}^k} = \sum_{i=1}^N |a_i^*| \|w_i^*\|_2^k$.

For $k \geq 1$, we apply the `torch.optim.Adam` optimizer (a variation on gradient descent), as implemented in PyTorch, to (9). We use a `betas` parameter $(0.9, 0.99)$, learning rate 0.5, number of iterations $T = 200$, penalty parameter $\lambda = 1$, and number of neurons $N = 10$. To enforce the nonnegativity condition on b, we project b to $[0, \infty)$ after each gradient step. Rather than take the last iterate, we choose the maximal MMD values among the iterates (after rescaling by the RTV^k seminorm of each iterate so that it lies in the unit seminorm ball). Further, we repeat this over three random initializations, and select the best resultant MMD value to be the final output.

For $k = 0$, such a first-order scheme is not applicable due to the fact that the gradient of the 0^{th} degree ridge spline $(w^\mathsf{T} x - b)_+^0 = 1\{w^\mathsf{T} x \geq b\}$ (with respect to w, b) is almost everywhere zero. As a surrogate, we directly approximate the optimum w^*, b^* in (2) using logistic regression.

4.2 Experimental setup

We compare the RKS tests of various degrees across several experimental settings, both to one another and to other all-purpose nonparametric tests: the energy distance test (Székely and Rizzo, 2005) as well as kernel MMD (Gretton et al., 2012) with a Gaussian kernel.

We fix the sample sizes to $m = n = 512$ throughout, and study four choices of dimension: $d = 2, 4, 8, 16$. For each dimension d, we consider five settings for P, Q, which are described in Table 1. In each setting, the parameter v controls the discrepancy between P and Q, but its precise meaning depends on the setting. The settings were broadly chosen in order to study the operating characteristics of the RKS test when differences between P and Q occur in one direction (settings 1–4), and in all directions (setting 5). Among the settings in which the differences occur in one direction, we also investigate different varieties (settings 1 and 2: mean shift under different geometries, setting 3: tail difference, setting 4: variance difference). Finally, we note that because the RKS test is *rotationally invariant*,[2] the fact that the chosen differences in Table 1 are axis-aligned is just a matter of convenience, and the results would not change if these differences instead occurred along arbitrary directions in $\mathbb{R}^d$.

For the RKS tests, we examine smoothness degrees $k = 0, 1, 2, 3$, and we center the input data to have the sample mean zero jointly across both samples. That is, each x_i is replaced by $x_j - \frac{1}{m+n} \big(\sum_{i=1}^m x_i + \sum_{i=1}^n y_i \big)$, and likewise for y_i. This makes the RKS test *translationally-invariant*. For kernel MMD, we set the Gaussian

[2] Meaning, $\rho(P_m, Q_n; \mathcal{F}_k)$ is invariant to rotations of the underlying samples $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^n$ by any arbitrary orthogonal transformation $U \in \mathbb{R}^{d \times d}$. This is true because the RTV^k seminorm is itself invariant to rotations of the underlying domain, as can be seen directly from the representation in Proposition 1.

Setting	P	Q	v
Pancake mean shift ("pancake-shift")	one axis $\sim \mathcal{N}(0,1)$ the others $\sim \mathcal{N}_{d-1}(0,16I)$	one axis $\sim \mathcal{N}(v,1)$ the others $\sim \mathcal{N}_{d-1}(0,16I)$	0.3
Ball mean shift ("ball-shift")	$\mathcal{N}_d(0,I)$	one axis $\sim \mathcal{N}(v,1)$ the others $\sim \mathcal{N}_{d-1}(0,I)$	0.2
t coordinate ("t-coord")	$\mathcal{N}_d(0,I)$	one axis $\sim t(v)$ the others $\sim \mathcal{N}_{d-1}(0,I)$	3
Variance inflation in one direction ("var-one")	$\mathcal{N}_d(0,I)$	one axis $\sim \mathcal{N}(0,v)$ the others $\sim \mathcal{N}_{d-1}(0,I)$	1.4
Variance inflation in all directions ("var-all")	$\mathcal{N}_d(0,I)$	$\mathcal{N}_d(0,vI)$	1.2

Table 1: Experimental settings. Here $\mathcal{N}(\mu,\Sigma)$ means the normal distribution with mean μ and covariance Σ, and $t(v)$ means the t distribution with v degrees of freedom.

kernel bandwidth according to a standard heuristic, based on the median of all pairwise squared distances between the input samples (Caputo et al., 2002). (The energy distance test has no tuning parameters.) For each setting, we compute these test statistics under the null where each x_i and y_i are sampled i.i.d. from the mixture $\frac{m}{m+n}P + \frac{n}{m+n}Q$, and under the alternative where x_i are i.i.d. from P and y_i from Q. We then repeat this 100 times (draws of samples, and computation of test statistics), and trace out ROC curves—true positive versus false positive rates—as we vary the rejection threshold for each test. Python code to replicate our experimental results is available at `https://github.com/100shpaik/`.

In Appendix K, we also compare the RKS test to kernel MMD with a polynomial kernel. An important remark is that the latter is not actually a nonparametric test, for any finite polynomial degree, because the population MMD does not define a metric (it cannot distinguish distributions P, Q that agree up to some number of moments, interpreted in the multivariate sense, but differ otherwise). We include comparisons to polynomial kernels for completeness nonetheless.

4.3 Experimental results

Figure 2 displays the ROC curves from all tests across all settings (rows) and all dimensions (columns). We see that in the first four settings (first four rows), the RKS tests exhibit generally strong performance: in each case, one of the k^{th} degree RKS tests performs either the best or near-best among the methods considered. This supports the intuition discussed earlier, since in each of these settings, P and Q differ in only a small number (in fact, one) direction. Conversely, in the last setting, "var-all", the distributions P and Q differ by a small amount in all directions, and the Gaussian kernel MMD and energy distance tests outperform the RKS tests, especially for larger d.

Keeping in mind that larger choices of k correspond to higher-order truncated moments, we can further inspect the first four settings to try to learn from why the different order RKS tests perform better or worse. In the "pancake-shift" setting, the mean shift occurs in a low-variance direction, leading to nearly separable classes, so the smallest choice $k = 0$ performs best, and better than kernel-based tests. In the "var-one" and "t-coord" settings, P and Q have the same mean, but different variance or tails, respectively, in one direction. Thus, larger choices of k perform better: $k = 1$ for "t-coord", and $k = 2, 3$ for "var-one". For a mean shift of between isotropic Gaussians, "ball-shift", degrees $k = 0, 1, 2$ are all reasonably competitive.

5 Discussion

We proposed and analyzed a new multivariate, nonparametric two-sample test, which is a maximum mean discrepancy (MMD) test with respect to a class of functions having unit Radon total variation, of a given

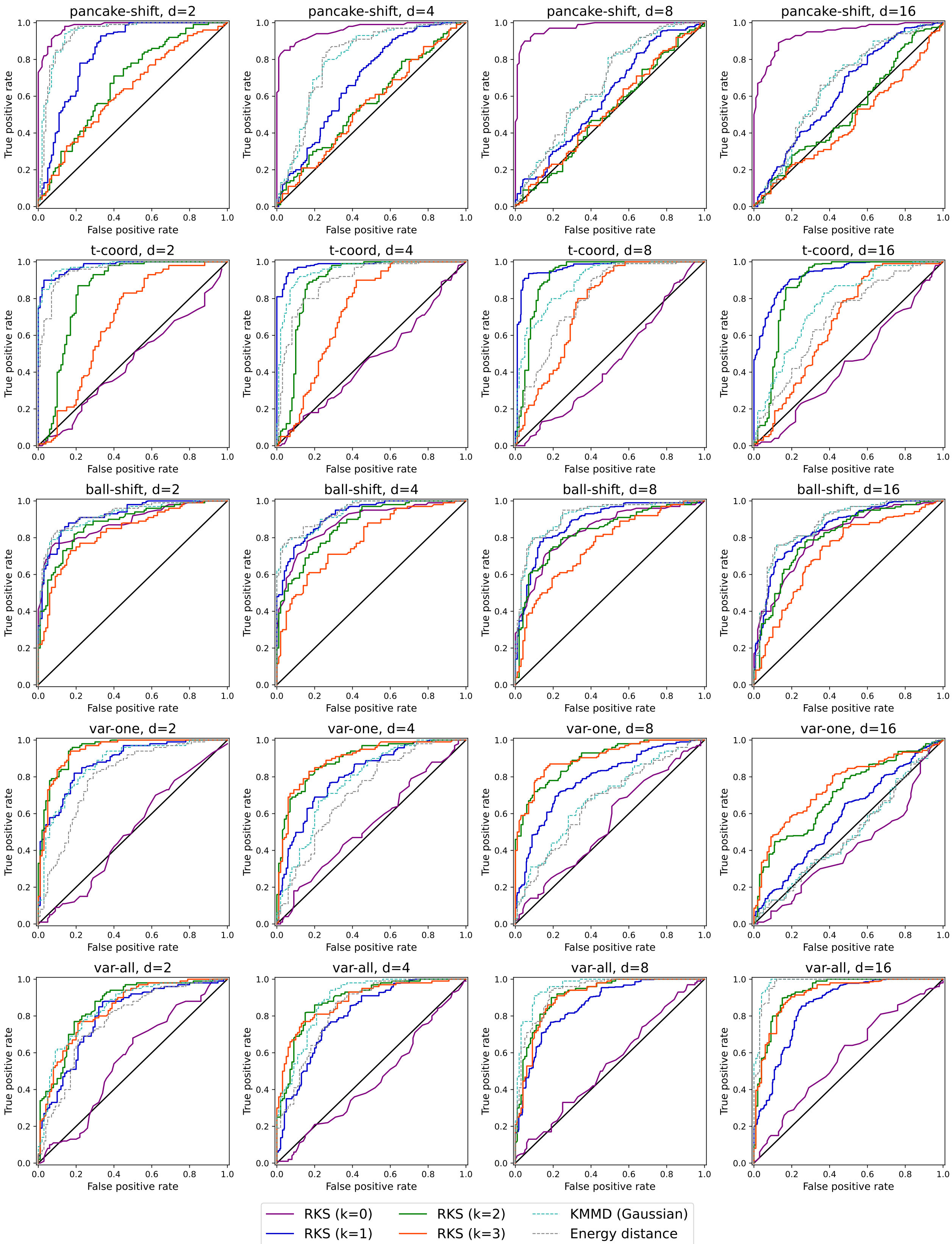


Figure 2: ROC curves across the experimental settings described in Table 1. Each row represents a different setting, and each column a different dimension. The RKS, kernel MMD (KMMD), and energy distance tests are compared (each coded by a combination of color and line type).

smoothness degree $k \geq 0$. We call this the Radon-Kolmogorov-Smirnov (RKS) test, and it generalizes the well-known Kolmogorov-Smirnov (KS) test to multiple dimensions and higher orders of smoothness.

The RKS test, like any statistical test, is not expected to be most powerful in all scenarios. Roughly, we expect the RKS test to be favorable in settings in which the given distributions P and Q differ in only a few directions in the ambient space; and higher orders of smoothness k can be more sensitive to tail differences along these few directions. When P and Q differ in many or all directions, the more traditional kernel MMD test with (say) Gaussian kernel will probably perform better.

It is worth revisiting computation of the RKS distance and discussing some limitations of our work and possible future directions. In this paper, we considered *full batch* gradient descent applied to (9) (we actually used Adam, but the differences for this discussion are unimportant), i.e., in each update the entire data set $\{x_i\}_{i=1}^m \cup \{y_i\}_{i=1}^n$ is used to calculate the gradient. This results in the following complexity: T iterations of gradient descent costs $O((n+m)dNT)$ operations, where recall N is the number of neurons. However, *mini batch* gradient descent should be also investigated, and is likely to be preferred in large problem sizes. That said, implementing mini batch gradients for (9) is nontrivial due to the nonseparable nature of the criterion across observations (due to the log and the absolute value in the first term), and pursuing this carefully is an important direction for future work. For now, we note that by just subsetting P_m, Q_n to a mini batch of size $b \leq n+m$, this results in a T iteration complexity of $O(bdNT)$ operations. To compare kernel MMD: this costs $O((n+m)^2 d)$ operations, and thus we can see that for large enough problem sizes computation of the RKS statistic via full batch and especially mini batch gradient descent can be more efficient, provided that small or moderate values of N, T still suffice to obtain good performance.

Finally, another direction worth pursuing is the use of deeper networks, which could be interpreted as learning the feature representation in which there is the largest discrepancy in projected truncated moments. This would help further distinguish the RKS test from kernel MMD, because the latter is "stuck with" the representation provided by the choice of kernel, and it cannot learn the representation based on the data. A deep RKS test could dovetail nicely with existing architectures (and even pre-trained networks) for related tasks in areas such as generative modeling, and out-of-distribution detection.

Acknowledgments

We thank Rahul Parhi and Jonathan Siegel and helpful discussions about Radon bounded variation spaces. This work was supported by NSF SCALE MoDL grant 2134133, ONR MURI grant N00014-20-1-2787, and the Miller Institute for Basic Research in Science at the University of California, Berkeley.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

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

Peter J. Bickel. A distribution free version of the Smirnov two sample test in the p-variate case. *Annals of Mathematical Statistics*, 40(1):1–23, 1969.

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

Barbara Caputo, K. Sim, Fredrik Furesjö, and Alexander J. Smola. Appearance-based object recognition using SVMs: Which kernel should i use? In *Advances in Neural Information Processing Systems Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision*, 2002.

Richard M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 2014. Second edition.

Ilya Dumer. Covering spheres with spheres. *Discrete & Computational Geometry*, 38:665–679, 2007.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. 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.

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

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

Youssef Mroueh, Chun Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. In *Proceedings of the International Conference on Learning Representations*, 2018.

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

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *Proceedings of the International Conference on Learning Representations*, 2020.

Rahul Parhi. *On Ridge Splines, Neural Networks, and Variational Problems in Radon-Domain BV Spaces*. PhD thesis, Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 2022.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

Rahul Parhi and Robert D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *Transactions on Information Theory*, 69(2):1125–1140, 2023.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, Aaditya Ramdas, and Ryan J. Tibshirani. A higher-order Kolmogorov-Smirnov test. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Proceedings of the Annual Conference on Learning Theory*, 2019.

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.

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.

Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

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

A Proof of Proposition 1

By Theorem 22 in Parhi and Nowak (2021), every $\mathsf{f} \in \mathrm{RBV}^k$ has a representative $f \in \mathsf{f}$ of the form

$$f(x) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \Big[(w^\mathsf{T} x - b)_+^k - \sum\nolimits_{|\alpha|\le k} c_\alpha(w,b)x^\alpha\Big] \mathrm{d}\mu(w,b) + q_0(x), \tag{10}$$

for a finite Radon measure μ on $\mathbb{S}^{d-1}\times\mathbb{R}$ such that: *(i)* μ is an odd (respectively even) measure if k is even (respectively odd); *(ii)* the integrand is integrable in (w,b) with respect to μ for any x; *(iii)* $\|\mu\|_{\mathrm{TV}} = \|f\|_{\mathrm{RTV}^k}$; *(iv)* $q_0(x)$ is a polynomial of degree at most k; *(v)* $\sum_{|\alpha|\le k}\big(c_\alpha(w,b)x^\alpha + (-1)^k c_\alpha(-w,-b)x^\alpha\big) = (w^\mathsf{T} x - b)^k$.

Using the conditions *(i)* and *(v)*, we can show the following with a manual calculation and the change of variables $(w,b)\to(-w,-b)$:

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}\times(-\infty,0)} \Big[(w^\mathsf{T} x - b)_+^k - \sum\nolimits_{|\alpha|\le k} c_\alpha(w,b)x^\alpha\Big]\, \mathrm{d}\mu(w,b)\\
&= \int_{\mathbb{S}^{d-1}\times(0,\infty)} \Big[(-w^\mathsf{T} x + b)_+^k - \sum\nolimits_{|\alpha|\le k} c_\alpha(-w,-b)x^\alpha\Big]\, \mathrm{d}\mu(-w,-b)\\
&= \int_{\mathbb{S}^{d-1}\times(0,\infty)} (-1)^{2k+1}\Big[(w^\mathsf{T} x - b)_-^k - \sum\nolimits_{|\alpha|\le k} (-1)^k c_\alpha(-w,-b)x^\alpha\Big]\, \mathrm{d}\mu(w,b)\\
&= \int_{\mathbb{S}^{d-1}\times(0,\infty)} (-1)^{2k+1}\Big[(w^\mathsf{T} x - b)_-^k - (w^\mathsf{T} x - b)^k + \sum\nolimits_{|\alpha|\le k} c_\alpha(w,b)x^\alpha\Big]\, \mathrm{d}\mu(w,b)\\
&= \int_{\mathbb{S}^{d-1}\times(0,\infty)} \Big[(w^\mathsf{T} x - b)_+^k - \sum\nolimits_{|\alpha|\le k} c_\alpha(w,b)x^\alpha\Big]\, \mathrm{d}\mu(w,b).
\end{aligned} \tag{11}$$

Define a measure $\tilde\mu$ by $\tilde\mu(E) = \mu(E_0) + 2\mu(E_+)$ where for an event E we denote $E_0 = (\mathbb{S}^{d-1}\times\{0\})\cap E$ and $E_+ = (\mathbb{S}^{d-1}\times(0,\infty))\cap E$. Then (11) rewrites (10) as

$$f(x) = \int_{\mathbb{S}^{d-1}\times[0,\infty)} \Big[(w^\mathsf{T} x - b)_+^k - \sum_{|\alpha|\le k} c_\alpha(w,b)x^\alpha\Big]\, \mathrm{d}\tilde\mu(w,b) + q_0(x) \tag{12}$$

where $\|\tilde\mu\|_{\mathrm{TV}} = \|\mu\|_{\mathrm{TV}} = \|f\|_{\mathrm{RTV}^k}$. Note that $(w^\mathsf{T} x - b)_+^k$ is integrable in (w,b) with respect to $\tilde\mu$ because $\|w\|_2 = 1$ and $b \ge 0$. Thus, $\int_{\mathbb{S}^{d-1}\times[0,\infty)} (w^\mathsf{T} x - b)_+^k \mathrm{d}\tilde\mu < \|x\|_2^k \|\tilde\mu\|_{\mathrm{TV}}$, hence $\sum_{|\alpha|\le k} c_\alpha(w,b)x^\alpha$ is integrable with respect to (w,b) for every x as well. Thus,

$$q_1(x) = \int_{\mathbb{S}^{d-1}\times[0,\infty)} \Big(\sum\nolimits_{|\alpha|\le k} c_\alpha(w,b)x^\alpha\Big) \mathrm{d}\tilde\mu(w,b)$$

is well-defined due to the condition *(ii)* and (12). Therefore (12) can be rewritten as

$$f(x) = \int_{\mathbb{S}^{d-1}\times[0,\infty)} (w^\mathsf{T} x - b)_+^k \,\mathrm{d}\tilde\mu(w,b) + \tilde{q_0}(x)$$

where $\tilde{q_0}(x) = q_0(x) + q_1(x)$ is a polynomial of degree at most k. Also $\tilde\mu|_{\mathbb{S}^{d-1}\times\{0\}} = \mu|_{\mathbb{S}^{d-1}\times\{0\}}$ by definition of $\tilde\mu$, and hence, $\tilde\mu|_{\mathbb{S}^{d-1}\times\{0\}}$ is odd when k is even, and even when k is odd by the condition *(i)*. Thus, we have arrived at the desired representation.

B Proof of Theorem 1

We rely on the representation established by Proposition 1 and treat the cases $k=0$ and $k\ge 1$ separately.

Case $k = 0$. For this case, the boundary condition (continuity at the origin) implies that the integral may be restricted to $b > 0$.

Lemma 1. *Consider* $\mathsf{f} \in \mathrm{RBV}^0$*. There exists a representative* $f \in \mathsf{f}$ *which is continuous at* $0 \in \mathbb{R}^d$ *if and only if* f *admits a representation*

$$f(x) = \int_{\mathbb{S}^{d-1}\times(0,\infty)} 1\{w^\mathsf{T} x \geq b\}\mathrm{d}\mu(w,b) + q, \tag{13}$$

where q *is a constant and* μ *is a signed measure with* $\|\mu\|_{\mathrm{TV}} < \infty$*.*

Proof of Lemma 1. By Proposition 1, there exists $f \in \mathsf{f}$ such that for all x,

$$f(x) = \int_{\mathbb{S}^{d-1}\times[0,\infty)} 1\{w^\mathsf{T} x \geq b\}\mathrm{d}\mu(w,b) + q = f^{=0}(x) + f^{>0}(x),$$

where

$$f^{>0}(x) = \int_{\mathbb{S}^{d-1}\times(0,\infty)} 1\{w^\mathsf{T} x \geq b\}\mathrm{d}\mu(w,b) + q,$$

$$f^{=0}(x) = \int_{\mathbb{S}^{d-1}\times\{0\}} 1\{w^\mathsf{T} x \geq 0\}\mathrm{d}\mu(w,b).$$

If $b > 0$, then $\lim_{x\to 0} 1\{w^\mathsf{T} x \geq b\} = 0$. Thus, by dominated convergence,

$$\lim_{x\to 0} f^{>0}(x) = q(x) = f^{>0}(0).$$

Thus, $f^{>0}$ is continuous at 0. We conclude that f is almost-everywhere equal to a function which is continuous at 0 if and only if $f^{=0}$ is almost-everywhere equal to a function which is continuous at 0.

For $x \neq 0$, $w^\mathsf{T} x \geq 0$ if and only if $w^\mathsf{T} x/\|x\| \geq 0$. Thus, $f^{=0}(x) = f^{=0}(x/\|x\|)$. Thus, $f^{=0}$ is almost everywhere equal to a function which is continuous at 0 if and only if there exists some $c \in \mathbb{R}$ such that for almost every $x \in \mathbb{S}^{d-1}$ (almost every with respect to Lebesgue surface measure), we have $f^{=0}(x) = c$. But this implies that $f^{=0}(x) = c$ for almost every $x \in \mathbb{R}^d$. Thus, we get that for almost every $x \in \mathbb{R}^d$,

$$f(x) = f^{>0}(x) + c = \int_{\mathbb{S}^{d-1}\times(0,\infty)} 1\{w^\mathsf{T} x \geq b\}\mathrm{d}\mu(w,b) + q + c.$$

The right-hand side is the desired representative. □

Theorem 1 in the case $k = 0$ is now a consequence of the following lemma, which additionally gives an explicit form for the radially cone continuous representative of f.

Lemma 2. *Assume* $\mathsf{f} \in \mathrm{RBV}^0$ *contains a representative which is continuous at* 0*. Then it contains a unique representative which is radially cone continuous. This representative can be written in the form* (13) *where* μ *is a signed measure with* $\|\mu\|_{\mathrm{TV}} < \infty$*. Conversely, any function of the form* (13) *where* μ *is a signed measure with* $\|\mu\|_{\mathrm{TV}} < \infty$ *is the unique radial cone continuous continuous representative of some* $\mathsf{f} \in \mathrm{RBV}^0$*.*

Proof of Lemma 2. Consider $\mathsf{f} \in \mathrm{RBV}^0$ which contains a representative which is continuous at the origin. By Lemma 1, there exists $f \in \mathsf{f}$ which satisfies (13) for all $x \in \mathbb{R}^d$, for some constant q and signed measure μ with $\|\mu\|_{\mathrm{TV}} < \infty$.

Fix $x \neq 0$. Consider any sequence $\epsilon_j \in \mathbb{R}$ such that $\epsilon_j \downarrow 0$, and $v_j \in \mathbb{S}^{d-1}$ such that $v_j \to x/\|x\|$. Consider any $(w,b) \in \mathbb{S}^{d-1}\times(0,\infty)$. If $w^\mathsf{T} x < b$, then $w^\mathsf{T}(x+\epsilon_j v_j) < b$ eventually because $x + \epsilon_j v_j \to x$. Alternatively, $w^\mathsf{T} x \geq b$ gives $w^\mathsf{T} x/\|x\| > 0$, whence $w^\mathsf{T} v_j > 0$ eventually. In particular, $w^\mathsf{T}(x+\epsilon_j v_j) > w^\mathsf{T} x > 0$ eventually. Thus, in both cases with have $1\{w^\mathsf{T}(x+\epsilon_j v_j) \geq b\} \to 1\{w^\mathsf{T} x \geq b\}$. By dominated convergence, we get that

$$\begin{aligned} f(x+\epsilon_j v_j) &= \int_{\mathbb{S}^{d-1}\times(0,\infty)} 1\{w^\mathsf{T}(x+\epsilon_j v_j) \geq b\}\mathrm{d}\mu(w,b) + q \\ &\to \int_{\mathbb{S}^{d-1}\times(0,\infty)} 1\{w^\mathsf{T} x \geq b\}\mathrm{d}\mu(w,b) + q = f(x). \end{aligned}$$

Thus, f is radially cone continuous and can be written in the form of (13).

By Theorem 22 in Parhi and Nowak (2021) and reversing the construction of $\tilde{\mu}$ from μ in the proof of Proposition 1, any function of the form (13) where μ is a signed measure with $\|\mu\|_{\mathrm{TV}} < \infty$ is a represenative of an equivalence class $\mathsf{f} \in \mathrm{RBV}^0$. The argument above also shows that it is radially cone continuous. Thus, we have shown the converse statement as well. □

Case $k = 1$**.** We begin with a lemma.

Lemma 3. *For $k \geq 1$ and every $\mathsf{f} \in \mathrm{RBV}^k$), there exists a unique representative $f \in \mathsf{f}$ which is continuous. This representative is in fact $(k-1)$-times continuously differentiable and can be written in the form in Proposition 1. Moreover, for any multi-index α with $|\alpha| \leq k-1$, we have for all $x \in \mathbb{R}^d$ that*

$$D^\alpha f(x) = \frac{k!}{(k-|\alpha|)!} \int_{\mathbb{S}^{d-1} \times [0,\infty)} w^\alpha (w^\mathsf{T} x - b)_+^{k-|\alpha|} \mathrm{d}\mu(w,b) + D^\alpha q(x). \tag{14}$$

where $w^\alpha = \prod_{i=1}^d w_i^{\alpha_i}$.

Proof of Lemma 3. For any $\mathsf{f} \in \mathrm{RBV}^k$, let $f \in \mathsf{f}$ be a representative of the form in Proposition 1. We will show that this representative f is $(k-1)$-times continuously differentiable by inducting on the order $m = |\alpha|$ of the multi-index α.

The base case is $m = 0$. In this case, the representation of the 0^{th} derivative is equivalent to the representation in the proposition. The continuity of f follows by Lemma 9 applied to this representation.

Now consider $m \geq 1$, and assume we have established the result for $m-1$. Consider α with $|\alpha| = m$, and without loss of generality, assume $\alpha_1 \geq 1$ and let $\tilde{\alpha} = (\alpha_1 - 1, \alpha_2, \ldots, \alpha_d)$. By the inductive hypothesis,

$$D^\alpha f(x) = \frac{k!}{(k-|\alpha|+1)!} \partial_{x_1} \int_{\mathbb{S}^{d-1} \times [0,\infty)} w^{\tilde{\alpha}} (w^\mathsf{T} x - b)_+^{k-|\alpha|+1} \mathrm{d}\mu(w,b) + D^\alpha q(x).$$

Note that for $x \in \mathbb{R}^d$, we have $\partial_{x_1} \big(w^{\tilde{\alpha}} (w^\mathsf{T} x - b)_+^{k-|\alpha|+1} \big) = (k - |\alpha| + 1) w^\alpha (w^\mathsf{T} x - b)^{k-|\alpha|}$. This function is integrable in (w,b) with respect to μ for every $x \in \mathbb{R}^d$, since $\|w\|_2 = 1$ and $b \geq 0$. Moreover, the partial derivative is uniformly bounded on $\{x' : \|x'\| \leq 2\|x\|\}$ by $2^{k-|\alpha|+1}(k-|\alpha|+1)\big(\prod_{i=1}^d |w_i|^{\alpha_i}\big)(2\|x\|)^{k-|\alpha|}$, which is integrable with respect to μ. These conditions allow us to apply the Leibniz rule to conclude that we can exchange differentiation and integration, which yields (14). The continuity of $D^\alpha f$ then follows from Lemma 9, and hence, the induction is complete.

Thus we have shown that f is $(k-1)$-times continuously differentiable. All that is left to show is that f is the unique element of f which is $(k-1)$-times continuously differentiable. This holds because any two continuous functions which are equal almost everywhere are in fact equal everywhere. $\square$

Theorem 1 in the case $k \geq 1$ is now a consequence of the following lemma, which additionally gives an explicit form for the radially cone continuous representative of f.

Lemma 4. *For $k \geq 1$, let $\mathsf{f} \in \mathrm{RBV}^k$ be such that the unique continuous representative $f \in \mathsf{f}$ is k-times classically differentiable at $0 \in \mathbb{R}^d$ with $D^\alpha f(0) = 0$ for all $|\alpha| \leq k$. Then f has representation of the form in Proposition 1 with $q(x) \equiv 0$.*

Proof of Lemma 4. Because $w^\alpha (w^\mathsf{T} 0 - b)_+^{k-|\alpha|} = 0$ for $|\alpha| \leq k-1$ and any $(w,b) \in \mathbb{S}^{d-1} \times [0,\infty)$, (14) implies that $D^\alpha f(0) = D^\alpha q(0)$ for all $|\alpha| \leq k-1$.

Next, we evaluate $D^\alpha f(0)$ for $|\alpha| = k$, which, by assumption, exists. Without loss of generality, assume $\alpha_1 \geq 1$, and let $\tilde{\alpha} = (\alpha_1 - 1, \alpha_2, \ldots, \alpha_d)$. Note that for $b > 0$, $\partial_{x_i} w^{\tilde{\alpha}} (w^\mathsf{T} x - b)_+ \big|_{x=0} = 0$. By (14), we can write

$$D^\alpha f(0) = k! \partial_{x_1} \int_{\mathbb{S}^{d-1} \times [0,\infty)} w^{\tilde{\alpha}} (w^\mathsf{T} x - b)_+ \,\mathrm{d}\mu(w,b) \Big|_{x=0} + D^\alpha q(0).$$

We compute the partial derivative in the first term manually:

$$\partial_{x_1} \int_{\mathbb{S}^{d-1} \times [0,\infty)} w^{\tilde{\alpha}} (w^\mathsf{T} x - b)_+ \,\mathrm{d}\mu(w,b) \Big|_{x=0} = \lim_{h \to 0} \frac{1}{h} \int_{\mathbb{S}^{d-1} \times [0,\infty)} w^{\tilde{\alpha}} (w_1 h - b)_+ \,\mathrm{d}\mu(w,b).$$

For $b > 0$ and $w \in \mathbb{S}^{d-1}$, we have $(w_1 h - b)_+ \leq |h| \cdot 1\{w_1 h > b\}$ because $|w_1| \leq 1$. Also $|w^\alpha| \leq 1$ because $|w_j| \leq 1$ for all j. Thus,

$$\begin{aligned} \left| \frac{1}{h} \int_{\mathbb{S}^{d-1} \times (0,\infty)} w^{\tilde{\alpha}} (w_1 h - b)_+ \,\mathrm{d}\mu(w,b) \right| &\leq \frac{1}{|h|} \int_{\mathbb{S}^{d-1} \times (0,\infty)} w^{\tilde{\alpha}} |h| 1\{w_j h - b\} \,\mathrm{d}\mu(w,b) \\ &\leq |\mu| \Big((w,b) \in \mathbb{S}^{d-1} \times (0,\infty) : w_j h > b \Big). \end{aligned}$$

Because $|\mu|$ is a finite positive measure and $\bigcap_{h>0}\big\{(w,b)\in\mathbb{S}^{d-1}\times(0,\infty): w_j h > b\big\}=\emptyset$, we conclude $|\mu|\big((w,b)\in\mathbb{S}^{d-1}\times(0,\infty): w_j h>b\big)\to 0$ as $h\to 0$. Thus, we have that

$$D^\alpha f(0) = k!\lim_{h\to 0}\Big[\frac{1}{h}\int_{\mathbb{S}^{d-1}\times\{0\}} w^{\tilde{\alpha}}(w_1h-b)_+\,\mathrm{d}\mu(w,b)\Big] + D^\alpha q(0), \tag{15}$$

and in particular, the limit in the first term exists. For $b=0$, $(w_1h-b)_+ = w_1h\cdot 1\{w_1>0\}$ when $h>0$, and $(w_1h-b)_+ = w_1h\cdot 1\{w_1<0\}$ when $h<0$. Thus, for $h>0$,

$$\frac{1}{h}\int_{\mathbb{S}^{d-1}\times\{0\}} w^{\tilde{\alpha}}(w_1h-b)_+\,\mathrm{d}\mu(w,b) = \int_{\substack{\mathbb{S}^{d-1}\times\{0\}\\ w_1>0}} w^\alpha\,\mathrm{d}\mu(w,b),$$

and for $h<0$,

$$\begin{aligned}\frac{1}{h}\int_{\mathbb{S}^{d-1}\times\{0\}} w^{\tilde{\alpha}}(w_1h-b)_+\,\mathrm{d}\mu(w,b) &= \int_{\substack{\mathbb{S}^{d-1}\times\{0\}\\ w_1<0}} w^\alpha\,\mathrm{d}\mu(w,b)\\ &= (-1)^{|\alpha|}(-1)^{k+1}\int_{\substack{\mathbb{S}^{d-1}\times\{0\}\\ w_1>0}} w^\alpha\,\mathrm{d}\mu(w,b),\end{aligned}$$

where the second equality uses the change of variables $w\mapsto -w$ and that $\mu|_{\mathbb{S}^{d-1}\times\{0\}}$ is odd when k is even and even when k is odd (see Proposition 1). By differentiability, the quantities in the two previous displays must be equal. Because $(-1)^{|\alpha|}(-1)^{k+1}=-1$, this implies that the quantities in the two previous displays must be 0. That is, the limit in (15) must be 0, and we have $0 = D^\alpha f(0) = D^\alpha q(0)$.

We have thus shown that $D^\alpha q(0)=0$ for all $|\alpha|\le k$, which, because $q(x)$ is a polynomial of degree at most k, implies $q(x)\equiv 0$. $\square$

C Proof of Theorem 2

Theorem 2 is an immediate consequence of Lemmas 2 and 4.

D Proof of Theorem 3

We again treat the cases $k=0$ and $k\ge 1$ separately.

Case $k=0$. Let

$$\mathcal{G}_0^+ = \Big\{1\{w^\top\cdot\ge b\}:\ (w,b)\in\mathbb{S}^{d-1}\times(0,\infty)\Big\}.$$

Because $\mathcal{G}_0^+\subseteq\mathcal{G}_0$, for any signed measure μ on $\mathbb{S}^{d-1}\times(0,\infty)$ with $\|\mu\|_{\mathrm{TV}}=1$ and $q\in\mathbb{R}$,

$$\begin{aligned}\sup_{f\in\mathcal{G}_0} P(f)-Q(f) \ge \sup_{f\in\mathcal{G}_0^+} P(f)-Q(f) &\ge \int_{\mathbb{S}^{d-1}\times(0,\infty)}\int_{\mathbb{R}^d} 1\{w^\top x\ge b\}\,\mathrm{d}(P-Q)(x)\,\mathrm{d}\mu(w,b)\\ &= \int_{\mathbb{R}^d}\Big(\int_{\mathbb{S}^{d-1}\times(0,\infty)} 1\{w^\top x\ge b\}\,\mathrm{d}\mu(w,b)+q\Big)\,\mathrm{d}(P-Q)(x).\end{aligned}$$

Taking the supremum on the right-hand side over measure μ, we get that

$$\sup\nolimits_{f\in\mathcal{G}_0} P(f)-Q(f)\ge \sup\nolimits_{f\in\mathcal{F}_0} P(f)-Q(f).$$

Because $\lim_{b\to\infty} P(w^\top x\ge b)-Q(w^\top x\ge b)$, $\mathbb{S}^{d-1}$ is compact, and $P(w^\top x\ge b)-Q(w^\top x\ge b)$ is continuous, the supremum on the left-hand side is achieved. Then, if the supremum is achieved at some (w,b) with $b>0$, since $x\mapsto 1\{w^\top x\ge b\}\in\mathcal{F}_0$, we also have

$$\sup_{f\in\mathcal{G}_0} P(f)-Q(f)\le \sup_{f\in\mathcal{F}_0} P(f)-Q(f).$$

On the other hand, consider that the supremum is achieved at some $(w, 0)$. Note that

$$\lim_{b \uparrow 0} Q(w^\top x \le b) - P(w^\top x \le b) = P(w^\top x \ge 0) - Q(w^\top x \ge 0).$$

But $Q(w^\top x \le b) - P(w^\top x \le b) = P(f)_b - Q(f)_b$ for $f_b(x) = 1\{-w^\top x \ge -b\}$. Note that $f_b \in \mathcal{G}_0^+$, and therefore, $f_b \in \mathcal{F}_0$. Thus, we have in this case also that $\sup_{f \in \mathcal{G}_0} P(f) - Q(f) \le \sup_{f \in \mathcal{F}_0} P(f) - Q(f)$.

Having shown both directions of the inequality, we conclude $\sup_{f \in \mathcal{G}_0} P(f) - Q(f) = \sup_{f \in \mathcal{F}_0} P(f) - Q(f)$. The same result holds with the roles of P and Q reversed, which establishes the result with the supremum over $|P(f) - P(Q)|$.

Case $k \ge 1$. Because P and Q have finite k^{th} moments, by dominated convergence

$$(w, b) \mapsto \mathbb{E}_P[(w^\top x - b)_+^k] - \mathbb{E}_Q[(w^\top x - b)_+^k]$$

is continuous on $\mathbb{S}^{d-1} \times [0, \infty)$. Thus,

$$\sup_{f \in \mathcal{G}_k} P(f) - Q(f) = \sup_{(w,b) \in \mathbb{S}^{d-1} \times (0,\infty)} \mathbb{E}_P[(w^\top x - b)_+^k] - \mathbb{E}_Q[(w^\top x - b)_+^k] \le \sup_{f \in \mathcal{F}_k} P(f) - Q(f),$$

where the last inequality holds because $x \mapsto (w^\top x - b)_+^k \in \mathcal{F}_k$ for all $w \in \mathbb{S}^{d-1}$ and $b > 0$ (Parhi and Nowak, 2021). On the other hand, for any $f \in \mathcal{F}_k$,

$$\begin{aligned}
P(f) - Q(f) &= \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1} \times [0,\infty)} (w^\top x - b)_+^k \, \mathrm{d}\mu(w, b) \, \mathrm{d}(P - Q)(x) \\
&= \int_{\mathbb{S}^{d-1} \times [0,\infty)} \int_{\mathbb{R}^d} (w^\top x - b)_+^k \, \mathrm{d}(P - Q)(x) \, \mathrm{d}\mu(w, b) \le \sup_{f' \in \mathcal{G}_k} |P(f') - Q(f')|,
\end{aligned}$$

where the second equality uses Fubini's theorem and the final inequality uses $\|\mu\|_{\mathrm{TV}} = \|\mathsf{f}\|_{\mathrm{RTV}^k} \le 1$. Because the previous display holds for all $f \in \mathcal{F}_k$, we have $\sup_{f \in \mathcal{F}_k} P(f) - Q(f) = \sup_{f \in \mathcal{G}_k} |P(f) - Q(f)|$. Since $f \in \mathcal{G}_k$ implies $-f \in \mathcal{G}_k$, we have $\sup_{f \in \mathcal{G}_k} |P(f) - Q(f)| = \sup_{f \in \mathcal{G}_k} P(f) - Q(f)$.

Thus, we have shown both that $\sup_{f \in \mathcal{G}_k} |P(f) - Q(f)| \le \sup_{f \in \mathcal{F}_k} |P(f) - Q(f)|$ and the reverse inequality, so that these are in fact equal.

E Proof of Theorem 4

If $P = Q$, then for any $f \in \mathcal{F}_k$, $P(f) - Q(f) = 0$, whence $\rho(P, Q; \mathcal{F}_k) = 0$. Alternatively, consider $P \ne Q$. Then, by the uniqueness of the characteristic function, there exists $w \in \mathbb{R}^d$ such that the distribution of $w^\top X$ and $w^\top Y$, where $X \sim P$ and $Y \sim Q$, are different. This implies that there exists some $t \in \mathbb{R}$ such that $P(w^\top X \ge t) \ne Q(w^\top Y \ge t)$.

In what follows, we treat the cases $k = 0$, $k = 1$, and $k \ge 2$ separately.

Case $k = 0$. If $P(w^\top X \ge t) > Q(w^\top Y \ge t)$ for some $t \ge 0$, then taking $f(x) = 1\{w^\top x \ge t\}$ gives

$$\rho(P, Q; \mathcal{F}_0) = \rho(P, Q; \mathcal{G}_0) \ge P(f) - Q(f) > 0,$$

where we have used Theorem 3. If $P(w^\top X \ge t) > Q(w^\top Y \ge t)$ for some $t \le 0$, then because $t \mapsto P(w^\top X \ge t)$ is left-continuous, for $s < t$ sufficiently close to t, we will have $P(w^\top X > s) > Q(w^\top Y > s)$. Then, taking $f(x) = -1\{(-w)^\top x \ge -s\}$ gives

$$\rho(P, Q; \mathcal{F}_0) = \rho(P, Q; \mathcal{G}_0) \ge P(f) - Q(f) > 0,$$

where we have used Theorem 3.

Similarly, if $P(w^\top X \ge t) < Q(w^\top X \ge t)$, then we can take $f(x) = -1\{w^\top x \ge t\}$ in the case $t \ge 0$ and $f(x) = 1\{(-w)^\top x \ge -s\}$ for some $s < t$ sufficiently close to t to conclude $\rho(P, Q; \mathcal{F}_0) > 0$.

Case $k = 1$**.** Because $\mathbb{E}_P[1\{w^\top X \geq s\}]$ and $\mathbb{E}_Q[1\{w^\top Y \geq s\}]$ are left-continuous in s, if $w^\top X$ and $w^\top y$ have a different distribution, there is some open interval $(\alpha, \beta) \subseteq \mathbb{R}$ on which $\mathbb{E}_P[1\{w^\top X \geq s\}] > \mathbb{E}_Q[1\{w^\top Y \geq s\}]$ or $\mathbb{E}_P[1\{w^\top X \geq s\}] < \mathbb{E}_Q[1\{w^\top Y \geq s\}]$, uniformly over all $s \in (\alpha, \beta)$. Because either $\beta > 0$ or $\alpha < 0$, this interval contains an open interval entirely contained in either $\mathbb{R}_{>0}$ or $\mathbb{R}_{<0}$. In the former case, we have that for $b > 0$, $\mathbb{E}_P[(w^\top X - b)_+] \neq \mathbb{E}_Q(w^\top Y - b)_+]$ because

$$\mathbb{E}_P\big[(w^\top X - b)_+\big] = \int_b^\infty \mathbb{E}_P\big[1\{w^\top X \geq s\}\big]\,\mathrm{d}s, \quad \mathbb{E}_Q\big[(w^\top Y - b)_+\big] = \int_b^\infty \mathbb{E}_Q\big[1\{w^\top Y \geq s\}\big]\,\mathrm{d}s.$$

We then get $\rho(P, Q; \mathcal{F}_1) = \rho(P, Q; \mathcal{G}_1) > 0$ by taking $f(x) = (w^\top x - b)_+$. In the latter case, we have that for $b > 0$, $\mathbb{E}_P[((-w)^\top X - b)_+] \neq \mathbb{E}_Q[((-w)^\top Y - b)_+]$ because

$$\mathbb{E}_P\big[((-w)^\top X - b)_+\big] = \int_{-\infty}^{-b} \mathbb{E}_P\big[(1 - 1\{w^\top X \geq s\})\big]\,\mathrm{d}s,$$

and likewise for Q. We then get $\rho(P, Q; \mathcal{F}_1) = \rho(P, Q; \mathcal{G}_1) > 0$ by taking $f(x) = ((-w)^\top x - b)_+$.

Case $k \geq 2$**.** Note that the derivatives of the mappings $b \mapsto \mathbb{E}_P[(w^\top X - b)_+^k]$ and $b \mapsto \mathbb{E}_Q[(w^\top Y - b)_+^k]$ are $b \mapsto k\mathbb{E}_P[(w^\top X - b)_+^{k-1}]$ and $b \mapsto k\mathbb{E}_Q[(w^\top Y - b)_+^{k-1}]$, respectively. If there is some (w, b) with $b \geq 0$ such that $\mathbb{E}_P[(w^\top X - b)_+^{k-1}] \neq \mathbb{E}_P[(w^\top X - b)_+^{k-1}]$, by continuity, $\mathbb{E}_P[(w^\top X - s)_+^{k-1}] \neq \mathbb{E}_P[(w^\top X - s)_+^{k-1}]$ for all s in an open interval around b, whence, by integration, $\mathbb{E}_P[(w^\top X - s)_+^k] \neq \mathbb{E}_P[(w^\top X - s)_+^k]$ for some s in this interval. We can then conclude the result by induction, using $k = 1$ as a base case.

F Proof of Theorem 5

For a function class $\mathcal{F}$, a probability measure $\mathbb{P}$, and $\epsilon > 0$, we say that $\{[l_i, u_i] : i \in [N]\}$ is an $(\epsilon, \mathbb{P})$-bracket of $\mathcal{F}$ with cardinality N if for all $i \in [N]$ we have $l_i(x) \leq u_i(x)$ for all x and

$$\|u_i(X) - l_i(X)\|_{L^2(\mathbb{P})}^2 = \mathbb{E}_{X \sim \mathbb{P}}[(u_i(X) - l_i(X))^2] \leq \epsilon^2,$$

and for all $f \in \mathcal{F}$ there exists $i \in [N]$ such that $l_i(x) \leq f(x) \leq u_i(x)$ for all x. The $(\epsilon, \mathbb{P})$-bracketing number, which we denote by $N_{[\,]}(\epsilon; \mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})})$, is the minimum cardinality of an $(\epsilon, \mathbb{P})$-bracket of $\mathcal{F}$. Using standard techniques from empirical process theory, we will be able to prove Theorem 5 by bounding the bracketing number $N_{[\,]}(\epsilon; \mathcal{G}_k, \|\cdot\|_{L^2(\mathbb{P})})$.

F.1 Bracketing number for $\mathcal{G}_k$ when $k \geq 1$

Theorem 7. *Let $k \geq 1$ and let $\mathbb{P}$ be a probability measure with finite $(2k + \Delta)$-moments for some $\Delta > 0$, i.e., $\mathbb{E}_{X \sim \mathbb{P}}\|X\|_2^{2k+\Delta} = M < \infty$. Then, for the class $\mathcal{G}_k$, there is a constant $C_{[\,]} > 0$ depending only on M, d, Δ, and k such that*

$$\log N_{[\,]}(\epsilon; \mathcal{G}_k, \|\cdot\|_2) \leq C_{[\,]} \log\Big(1 + \frac{1}{\epsilon}\Big).$$

Proof of Theorem 7. We denote by $\mathsf{B}_r^d(x)$ the ball of radius r centered at $x \in \mathbb{R}^d$:

$$\mathsf{B}_r^d(x) = \{z \in \mathbb{R}^d \mid \|z - x\|_2^2 \leq r^2\}.$$

We say a set $S \subseteq \mathbb{R}^d$ is a r-covering of a set $A \subseteq \mathbb{R}^d$ if A is contained in the union of the balls of radius ϵ centered at points $x \in I$:

$$A \subseteq \bigcup\nolimits_{x \in I} \mathsf{B}_r^d(x).$$

For a compact set $I \subseteq \mathbb{R}^d$ and real number $b \in \mathbb{R}$, define functions

$$\ell_{I,b}(x) = \inf\Big\{(w^\top x - b)_+^k : w \in I\Big\}, \quad \text{and} \quad u_{I,j}(x) = \sup\Big\{(w^\top x - b)_+^k : w \in I\Big\},$$

where we adopt the convention that $\ell_{I,\infty}(x) = u_{I,\infty}(x) \equiv 0$.

Lemma 5. *Consider S is an r-covering of $\mathbb{S}^{d-1}$. Let $\mathcal{I}$ be the collection of sets $\{I = \mathbb{S}^{d-1} \cap \mathsf{B}^d_r(x) : x \in S\}$, and let $b_j = j\delta$ for $j = -1, \ldots, N$, and $b_{N+1} = \infty$ for some $\delta > 0$ and integer $N > 0$. Then, the following is a bracket of $\mathcal{G}_k$*

$$\big\{[l_{I,b_{j+1}}, u_{I,b_j}] : I \in \mathcal{I}, j = -1, \ldots, N\big\}. \tag{16}$$

Proof of Lemma 5. Consider any $(w, b) \in \mathbb{S}^{d-1} \times [0, \infty)$ and its corresponding function $x \mapsto (w^\top x - b)^k_+ \in \mathcal{G}_k$. Because S is an r-covering, there exists $I \in \mathcal{I}$ such that $w \in I$. Also clearly there exists $j \in \{-1, 0, 1, \ldots, N\}$ such that $b_j < b \le b_{j+1}$. Then $\ell_{I,j} \le (w^\top x - b)^k_+$:

$$\ell_{I,b_j}(x) = \inf\Big\{(\tilde{w}^\top x - b_j)^k_+ : w \in I\Big\} \le (w^\top x - b_{j+1})^k_+ \le (w^\top x - b)^k_+.$$

Similarly, we can show $(w^\top x - b)^k_+ \le u_{I,j}(x)$. □

Lemma 6. *In the setting of Lemma 5, if we set $r = \delta = \frac{\epsilon/3}{k\sqrt{1+\max(1,M)}}$ and $N = \lceil (\frac{M}{\epsilon})^{1/\Delta} \frac{1}{\delta} \rceil$, then we have $\|u_{I,b_j} - l_{I,b_{j+1}}\|_{L^2(P)} \le \epsilon$ for all $j = -1, \ldots, N$ and $I \in \mathcal{I}$.*

Proof of Lemma 6. We bound

$$\begin{aligned}
|u_{I,b_j}(x) - l_{I,b_{j+1}}(x)| &\le \max_{w,v \in I} |(w^\top x - b_j)^k_+ - (v^\top x - b_{j+1})^k_+| \\
&\le \max_{w,v \in I} |(w^\top x - b_j)^k_+ - (v^\top x - b_j)^k_+| + \max_{v \in I} |(v^\top x - b_j)^k_+ - (v^\top x - b_{j+1})^k_+|.
\end{aligned}$$

Using a telescoping sum, we may bound for $j \le N$

$$\begin{aligned}
|(w^\top x - b_j)^k_+ - (v^\top x - b_j)^k_+| &= |(w^\top x - b_j)_+ - (v^\top x - b_j)_+| \times \sum\nolimits_{i=0}^{k-1} (w^\top x - b_j)^i_+ (v^\top x - b_j)^{k-1-i}_+ \\
&\le \|w - v\|_2 \|x\|_2 \times k \cdot \max\{(w^\top x)_+, (v^T x)_+\}^{k-1} \\
&\le k\|w - v\|_2 \|x\|_2^k \\
&\le 2kr\|x\|_2^k.
\end{aligned}$$

Likewise, we may bound for $j \le N - 1$

$$\begin{aligned}
|(v^\top x - b_j)^k_+ - (v^\top x - b_{j+1})^k_+| &= |(v^\top x - b_j)_+ - (v^\top x - b_{j+1})_+| \times \sum\nolimits_{i=0}^{k-1} (v^\top x - b_j)^i_+ (v^\top x - b_{j+1})^{k-1-i}_+ \\
&\le |b_j - b_{j+1}| \times k \cdot (v^\top x - b_j)^{k-1-i}_+ \\
&\le k\delta \|x\|_2^{k-1}.
\end{aligned}$$

Therefore, for $j \le N - 1$,

$$|u_{I,b_j}(x) - l_{I,b_{j+1}}(x)| \le k\delta\|x\|_2^{k-1} + 2kr\|x\|_2^k \le \begin{cases} k(\delta + 2r) & : \|x\|_2 \le 1 \\ k(\delta + 2r)\|x\|_2^k & : \|x\|_2 > 1 \end{cases},$$

which can be simplified to

$$|u_{I,b_j}(x) - l_{I,b_{j+1}}(x)| \le k(\delta + 2r) \max\{1, \|x\|_2^k\}.$$

This gives for $j \le N - 1$,

$$\|u_{I,b_j} - l_{I,b_{j+1}}\|^2_{L^2(\mathbb{P})} \le k^2(\delta + 2r)^2 (1 + \mathbb{E}_{X \sim P}\|X\|_2^{2k}) \le k^2(\delta + 2r)^2 \cdot \max(1, M) \le \epsilon^2,$$

where the final inequality is obtained by plugging in the values of r and δ specified in the statement of the lemma.

Meanwhile, for $j = N$, we have $b_{j+1} = \infty$, so that $l_{I,b_{j+1}} = 0$. Also, $b_j = N\delta \geq (M/\epsilon)^{1/\Delta}$. Thus,

$$\begin{aligned}
\|u_{I,b_N} - l_{I,b_{N+1}}\|^2_{L^2(\mathbb{P})} &= \int_{\mathbb{R}^d} u_{I,b_N}(x)^2 \mathrm{d}\mathbb{P}(x) \\
&\leq \int_{\mathbb{R}^d} (\|x\|_2 - b_N)_+^{2k} \mathrm{d}\mathbb{P}(x) = \int_{\mathbb{R}^d} (\|x\|_2 - N\delta)_+^{2k} \mathrm{d}\mathbb{P}(x) \\
&\leq \Big(\int_{\mathbb{R}^d} (\|x\|_2 - N\delta)^{2k+p} \mathrm{d}\mathbb{P}(x) \Big)^{\frac{2k}{2k+\Delta}} \Big(\int_{\mathbb{R}^d} 1(\|x\|_2 \geq N\delta)\ \mathrm{d}\mathbb{P}(x) \Big)^{\frac{\Delta}{2k+\Delta}} \\
&= M^{\frac{2k}{2k+\Delta}} \times \mathbb{P}\Big(\|X\|_2 \geq C\delta \Big)^{\frac{p}{2k+\Delta}} \\
&\leq M^{\frac{2k}{2k+\Delta}} \times \Big(\frac{M}{(N\delta)^{2k+\Delta}} \Big)^{\frac{\Delta}{2k+\Delta}} = \frac{M}{(N\delta)^{\Delta}},
\end{aligned}$$

where the second inequality uses Hölder's inequality, and the last inequality used Markov inequality. With the value of δ and N specified by the lemma, we conclude $\|u_{I,b_N} - l_{I,b_{N+1}}\|^2_{L^2(\mathbb{P})} \leq \epsilon^2$. $\square$

We can now complete the proof of Theorem 7. By Dumer (2007), there exists an r-covering of $\mathbb{S}^{d-1}$ with cardinality at most $(1 + 2/r)^d$. Thus, using the choices of r, δ, and N in Lemma 6 gives an $(\epsilon, \mathbb{P})$-bracket of $\mathcal{G}_k$ that has cardinality at most

$$\Big(1 + \frac{2k\sqrt{1 + \max(1, M)}}{\epsilon/3} \Big)^d \Big((\frac{M}{\epsilon})^{1/\Delta} \frac{k\sqrt{1 + \max(1, M)}}{\epsilon/3} + 1 \Big) \leq C(1 + 1/\epsilon)^{d+1+1/\Delta}.$$

with some constant C. The proof is completed. $\square$

F.2 Bracketing number for $\mathcal{G}_0$

Theorem 8. *Let $\mathbb{P}$ be a probability measure with finite Δ-moments for some $\Delta > 0$, i.e., $\mathbb{E}_{X\sim\mathbb{P}}\|X\|_2^{\Delta} = M < \infty$. Suppose additionally that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure and has bounded density, i.e. $\sup_{x\in\mathbb{R}^d} |p(x)| \leq M_\infty < \infty$. hen, for the class $\mathcal{G}_0$, there is a constant $C_{[\,]} > 0$ depending only on M, M_∞, d, and Δ such that*

$$\log N_{[\,]}(\epsilon; \mathcal{G}_0, \|\cdot\|_2) \leq C_{[\,]} \log \Big(1 + \frac{1}{\epsilon} \Big).$$

Proof. We adopt the same notation used in the proof of Theorem 7. The set (16) is a bracket of $\mathcal{G}_0$ by Lemma 5. Now we show that it is an ϵ-bracket for appropriate choices of r, δ, and N.

Select $I \in \mathcal{I}, j = -1, \ldots, N-1$. There exists a $w_0 \in I$ such that $\|w - w_0\|_2 \leq r$ for all $w \in I$, since S is an r-cover of $\mathbb{S}^{d-1}$. Consequently, for any $x \in \mathbb{R}^d$

$$\begin{aligned}
\max_{w\in I}\ 1(w^\top x \geq b_j) \ &\leq\ 1(w_0^\top x \geq b_j - r\|x\|_2), \\
\min_{v\in I}\ 1(v^\top x \geq b_{j+1}) \ &\geq\ 1(w_0^\top x \geq r\|x\|_2 + b_{j+1})
\end{aligned}$$

Notice that since $w_0 \in \mathbb{S}^{d-1}$, the random variable $w_0^\top X \in \mathbb{R}$ has a probability density that is upper bounded by M_∞. Therefore, for any $R > 0$,

$$\begin{aligned}
\|u_{I,b_j} - l_{I,b_j}\|^2_{L^2(\mathbb{P})} &= \mathbb{P}\Big(\max_{w,v} \Big| 1(w^\top x \geq b_j) - \min_{v\in I}\ 1(v^\top x \geq b_{j+1}) \Big| \Big) \\
&\leq \mathbb{P}\Big(w_0^\top X \in [b_j - r\|X\|_2, b_{j+1} + r\|X\|_2] \Big) \\
&\leq \mathbb{P}\Big(w_0^\top X \in [b_j - rR, b_{j+1} + rR] \Big) + \mathbb{P}\Big(\|X\|_2 \geq R \Big) \\
&\leq M_\infty \Big(\delta + 2rR \Big) + \mathbb{P}\Big(\|X\|_2 \geq R \Big) \\
&\leq M_\infty \Big(\delta + 2rR \Big) + \frac{M}{R^\Delta},
\end{aligned}$$

with the last line following from Markov's inequality. Taking $R = (3M/\epsilon^2)^{1/\Delta}$, $r = \frac{\epsilon^2}{6RM_\infty} = \frac{\epsilon^{2(1+1/\Delta)}}{6(3M)^{1/\Delta}M_\infty}$ and $\delta = \frac{\epsilon^2}{3M_\infty}$ makes the above upper bound equal to ϵ^2.

On the other hand for $I \in \mathcal{I}, j = N$, we have $l_{w,b_N+1} = 0$, and

$$\begin{aligned}\|u_{I,b_N} - l_{I,b_{N+1}}\|_P^2 &= \mathbb{P}\Big(\max_{w\in I} w^\top X \geq b_N\Big) \\ &\leq \mathbb{P}\Big(\|X\|_2 \geq b_N\Big) \\ &\leq \frac{M}{(N\delta)^\Delta}.\end{aligned}$$

Taking $N = \frac{(M/\epsilon^2)^{1/\Delta}}{\delta}$ makes this upper bound equal to ϵ^2. So for these choices of r, δ and N the bracket is an ϵ-bracket.

Finally, arguing as in the proof of Theorem 7 we conclude that there exists an ϵ-bracket of $\mathcal{G}_0$ with cardinality at most

$$N \times \Big(1 + \frac{2}{r}\Big)^d \leq C(1 + \frac{1}{\epsilon})^{2d(1+1/\Delta)+2(1+1/\Delta)}.$$

This completes the proof. □

F.3 Proof of Theorem 5

Theorem 5 is a consequence of the $(\epsilon, \mathbb{P})$-bracketing number bound of Theorem 7 (and when $k = 0$, Theorem 8) and the following result from empirical process theory, which we copy for convenience.

Theorem 9 (Theorems 7.6 and 9.1 in Dudley (2014))**.** *Let (Ω, Σ, P) be a probability space and let $\mathcal{F} \subseteq \mathcal{L}^2(\Omega, \Sigma, P)$ be such that*

$$J_{[\,]}(\mathcal{F}, \|\cdot\|_2) = \int_0^1 \sqrt{\log N_{[\,]}(\epsilon; \mathcal{F}, \|\cdot\|_{L^2(P)})}\, \mathrm{d}\epsilon < \infty.$$

Then, under the P-null hypothesis, as $m, n \to \infty$,

$$\sqrt{\frac{1}{m} + \frac{1}{n}}\,(P_m - Q_n) \xrightarrow{d} \mathbb{G}_P, \quad \textit{and hence} \quad \sqrt{\frac{1}{m} + \frac{1}{n}}\, \sup_{f\in\mathcal{F}} \big|P_m f - Q_n f\big| \xrightarrow{d} \sup_{f\in\mathcal{F}} |\mathbb{G}_P(f)|.$$

Theorem 9 implies Theorem 5 because $\int_0^1 \sqrt{\log(1 + 1/\epsilon)}\, \mathrm{d}\epsilon < \infty$.

G Proof of Theorem 6

Our proof follows closely the proof of the analogous result for the univariate higher-order Kolmogorov-Smirnov test in Sadhanala et al. (2019). We begin by recalling the following lemma, which appears as Theorem 2.14.2 and 2.14.5 in van der Vaart and Wellner (1996), the statement of which we transcribe from Sadhanala et al. (2019).

Lemma 7. *Let $\mathcal{F}$ be a class of functions with an envelope function F; i.e., $f \leq F$ for all $f \in \mathcal{F}$. Define*

$$W = \sqrt{n}\sup_{f\in\mathcal{F}} |P_n f - Pf|,$$

and let $J = \int_0^1 \sqrt{\log N_{[]}(\epsilon; \mathcal{F}, \|\cdot\|_{L^2(P)})}\mathrm{d}\epsilon$. If for $p \geq 2$, $\|F\|_{L^p(P)} < \infty$, then for a universal constant $c_1 > 0$,

$$\mathbb{E}[|W|^p]^{1/p} \leq c_1\big(\|F\|_{L^2(P)} J + n^{-1/2+1/p}\|F\|_{L^p(P)}\big).$$

If, for $0 < p \leq 1$, the exponential Orlicz norm of F, i.e. $\|F\|_{\Psi_p} = \inf\{t > 0 : \mathbb{E}[\exp(|F|^p/t^p)] \leq 2\}$, is finite, then

$$\|W\|_{\Psi_p} \leq c\big(\|F\|_{L^2(P)} J + n^{-1/2}(1 + \log n)^{1/2}\|F\|_{\Psi_p}\big).$$

Using this lemma, we have the following theorem on the concentration of the the RKS test statistic on its population value.

Theorem 10 (Two sided tail bound for $T_{d,k}$)**.** *We have the following tail bounds.*

(i) Consider $k = 0$. Assume P and Q have Δ^{th} moment bounded by M for $M, \Delta > 0$, which are absolutely continuous with respect to Lebesuge measure, and have densities bounded by M_∞. Then, for some constant c_0 depending only on M, Δ, M_∞, and d,

$$|T_{d,k} - \rho(P, Q; \mathcal{F}_k)| \le c_0(\log(1/\alpha))\Big(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\Big).$$

(ii) Consider $k \ge 1$. Assume P and Q have finite p^{th} moments bounded by M for $p = 2k + \Delta$ for $M, \Delta > 0$, then there exists a constant c_0 depending only on M, Δ, k, and d, such that

$$|T_{d,k} - \rho(P, Q; \mathcal{F}_k)| \le \frac{c_0}{\alpha^{1/p}}\Big(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\Big)$$

with probability at least $1 - \alpha$. Assume, on the other hand, that $\|X\|$ and $\|Y\|$ have Orlicz norm of order $0 < p \le 1$ bounded by M when $X \sim P$ and $Y \sim Q$. Then there exists a constant c_0 depending only on M, p, k, and d, such that

$$|T_{d,k} - \rho(P, Q; \mathcal{F}_k)| \le c_0(\log(1/\alpha))^{1/p}\Big(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\Big).$$

Proof of Theorem 10. First consider $k \ge 1$. Note that a bound on any Orlicz-norm of order $0 < p \le q$ implies a bound on the $(2k + \Delta)^{\text{th}}$-moments of P, Q for $\Delta = 1$. Thus, assuming either the moment bound or the Orlicz norm bound, Theorem 7 implies $J = \int_0^1 \sqrt{\log N_{[]}(\epsilon; \mathcal{G}_k, \|\cdot\|_{L^2(P)})}$ has a finite upper bound depending only on the $(2k+\Delta)^{\text{th}}$-moments of P and Q, d, Δ, and k. Take $F(x) = \|x\|_2^k$. If we only assume a bound on the $(2k+\Delta)^{\text{th}}$ moments of P and Q, then take $p = 2k + \Delta \ge 2$. In this case, $n^{-1/2+1/p} \le 1$, whence using Lemma 7 and Markov's inequality, we see that $\sup_{f\in\mathcal{F}} |P_n f - Pf| \le c_0/(\alpha^{1/p}\sqrt{n})$ and $\sup_{f\in\mathcal{F}} |Q_n f - Qf| \le c_0/(\alpha^{1/p}\sqrt{m})$ with probability at least $1-\alpha$. The asserted bound then holds by the triangle inequality. On the other hand, if we assume a bound on the Orlicz norm of order p, Lemma 7 implies a constant upper bound on the Orlicz norm of order p of $\sqrt{n}\sup_{f\in\mathcal{F}} |P_n f - Pf|$ and $\sqrt{m}\sup_{f\in\mathcal{F}} |Q_n f - Qf|$ because $(1 + \log n)^{1/2}/n^{1/2} \le 1$. This gives us the desired tail bound by taking $t = c_0(\log(1/\alpha))^{1/p}$ in the supremum defining the Orlicz norm and Markov's inequality.

Now consider $k = 0$. We take F as a constant 1 function, i.e. $F = 1$, and using Theorem 8 to bound the integral J in this case. Because $F = 1$ has a finite Orlicz norm of order $p = 1$, the asserted tail bound follows from Lemma 7. ☐

We can now complete the proof of Theorem 6. By Theorem 5, if $P = Q$, then $T_{d,k} = O_p(\sqrt{(n+m)/(nm)}$. Because $1/t_{n+m} = o(\sqrt{n+m}) = o(\sqrt{(nm)/(n+m)})$, we see that the test rejects with asymptotic probability 0. On the other hand, if $P \ne Q$, by Theorem 4, $\rho(P, Q; \mathcal{F}_k) > 0$. By Theorem 10, $T_{d,k} \xrightarrow{p} \rho(P, Q; \mathcal{F}_k)$. Because $t_{n+m} \to 0$, this implies that the test rejects with asymptotic probability 1. The proof is complete.

H Auxilliary technical lemmas

Lemma 8. *For $a, b \in \mathbb{R}$ and integer $k \ge 1$,*

$$\big|a_+^k - b_+^k\big| \le k(a_+^{k-1} + b_+^{k-1})|a - b|,$$

with the convention that when $k - 1 = 0$, we set $0^0 = 1$.

Proof of Lemma 8. Without loss of generality, assume $a \le b$. The function $x \mapsto x_+^k$ has derivative kx_+^{k-1} almost everywhere (even under our convention for when $k - 1 = 0$). For every point $x \in [a, b]$, we have $|kx_+^{k-1}| \le k(a_+^{k-1} + b_+^{k-1})$, because $|x_+| \le \max\{a_+, b_+\}$. Thus, integrating the derivative along the path between a and b gives the result. ☐

Lemma 9. *Assume $\mu \in \mathcal{M}(\mathbb{S}^{d-1} \times [0,\infty))$. Then, for any $0 \le m \le k-1$ and any function $c(w,b) : \mathbb{S}^{d-1} \times [0,\infty) \to \mathbb{R}$ which is bounded by $|c(w,b)| \le C$ for some constant C, we have*

$$x \mapsto \int_{\mathbb{S}^{d-1}\times[0,\infty)} \Big(c(w,b)(w^\top x - b)_+^{k-m} \Big) \, \mathrm{d}\mu(w,b) \tag{17}$$

is continuous at every $x \in \mathbb{R}^d$.

Proof of Lemma 9. Denote $f(x) = \int_{\mathbb{S}^{d-1}\times[0,\infty)} c(w,b)(w^\top x - b)_+^{k-m} \, \mathrm{d}\mu(w,b)$. Let $|\mu|$ be the total variation measure of the signed measure μ. Note that $\|\mu\|_{\mathrm{TV}} = \| \, |\mu| \, \|_{\mathrm{TV}}$. For $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
\big|f(x) - f(y)\big| &\le \int_{\mathbb{S}^{d-1}\times[0,\infty)} |c(w,b)| \, \big|(w^\top x - b)_+^{k-m} - (w^\top y - b)_+^{k-m}\big| \, \mathrm{d}|\mu|(w,b) \\
&\le (k-m) \int_{\mathbb{S}^{d-1}\times[0,\infty)} |c(w,b)| \, \big((w^\top x - b)_+^{k-m-1} + (w^\top y - b)_+^{k-m-1}\big) |w^\top (x-y)| \, \mathrm{d}|\mu|(w,b) \\
&\le (k-m)\|x-y\| \int_{\mathbb{S}^{d-1}\times[0,\infty)} |c(w,b)| \, \big((\|w\|_2 \|x\|_2)^{k-m-1} + (\|w\|_2 \|y\|_2)^{k-m-1}\big) \, \mathrm{d}\mu(w,b) \\
&\le (k-m)\|x-y\| \, C \, \|\mu\|_{\mathrm{TV}} \big(\|x\|_2^{k-m-1} + \|y\|_2^{k-m-1}\big),
\end{aligned}$$

where in the second inequality we have used Lemma 8. Since the right-hand side goes to 0 as $y \to x$, the proof is complete. □

I Sensitivity analysis: log transform

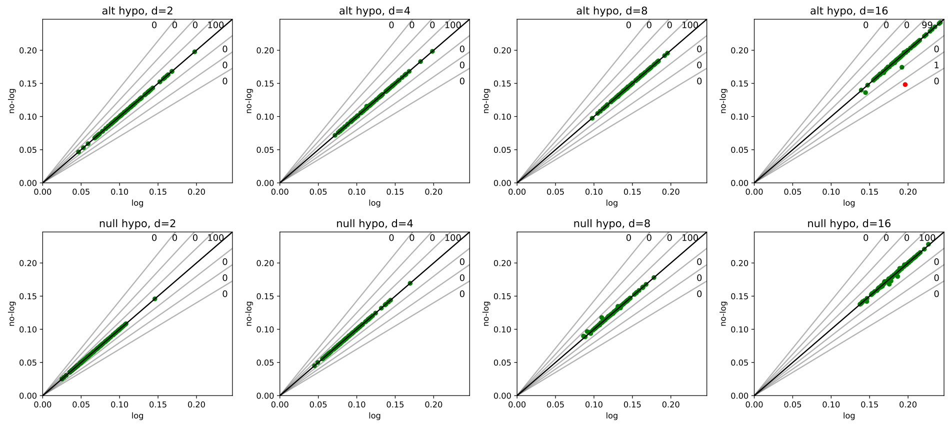
In (9), which is the basis for the experiments in the main paper, we use a log transform of the MMD term in the criterion. The current section compares the performance of (local) optimizers of this problem, which we call the "log" problem, for short, to those of

$$\min_{\substack{f = f_{(a_j, w_j, b_j)_{j=1}^N} \\ b_j \ge 0, \, j=1,\dots,N}} \; -\frac{1}{kN} \big| P_m(f) - Q_n(f) \big| + \frac{\lambda}{k} \bigg(\frac{1}{N} \sum_{i=1}^N |a_i| \|w_i\|_2^k \bigg)^2, \tag{18}$$

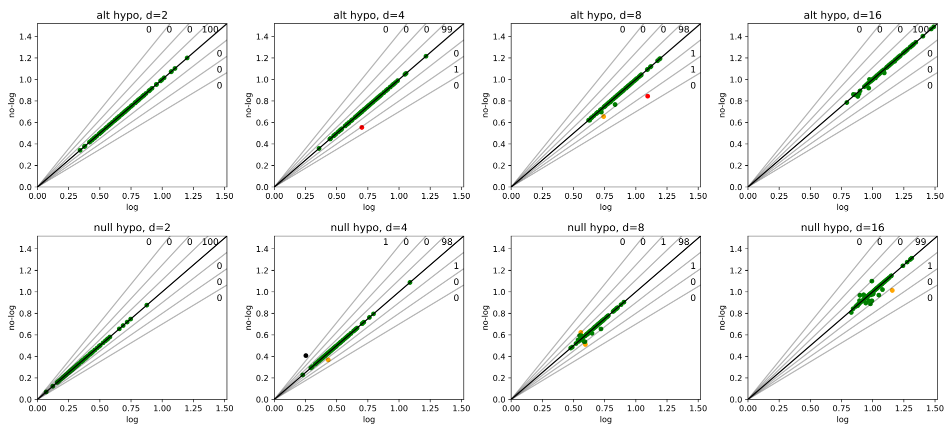
which we call the "no-log" problem, for short. In the above display, we can see that there is no log transform on the MMD term, but in addition, the penalty on the RTV^k seminorm has been squared. This is done because without such a transformation, one can show that the problem (MMD term plus RTV^k seminorm) does not attain its infimum, unless λ is chosen very carefully so that the infimum in zero. By squaring the penalty, the criterion in (18) is coercive, which guarantees (since it is also continuous) that it will attain its infimum. In both (9) and (18), we fix $\lambda = 1$, and use `torch.optim.Adam` with `betas` parameter $(0.9, 0.99)$. We explore the behavior for different learning rates and numbers of iterations. Throughout, we focus on the "var-one" setting described in Table 1, and use $N = 10$ neurons. The summary of our findings is as follows.

- Optimizing the "log" problem typically results in a larger MMD value compared to the "no-log" problem, especially for larger k, d, and especially under the null.
- Optimizing the "log" and "no-log" problems generally results in similar ROC curves, however, for larger k, d, the "no-log" ROC curves can actually be slightly better. This actually appears to be driven by the fact that "no-log" optimization is *relatively worse* in terms of the MMD values obtained under the null, which gives it a slightly greater separation and slightly higher power.
- The "log" problem is more robust to the choice of learning rate, both in terms of the MMD values and ROC curves obtained, especially for larger k, d.
- The "log" problem typically shows faster convergence (number of iterations required to approach a local optimum), especially for larger k, d.

Figures 3–9 display the results from our sensitivity analyses. In each one, the figure caption explains the salient points about the setup and takeaways. We remark that results for $k = 4$ (not included for brevity) show qualitatively similar but even more extreme behavior.



(a) $k = 1$



(b) $k = 3$

Figure 3: MMD values obtained by optimizing the "log" (x-axis) and "no-log" (y-axis) problems. Each point represents a set of samples drawn from P, Q, and the result of running $T = 1200$ iterations with learning rate 0.01, for each criterion. (This learning rate was chosen to be favorable to the "no-log" problem.) Points below the diagonal mean that the "log" criterion results in a larger MMD value, which we see is especially prominent for larger k, d, and more prominent under the null.

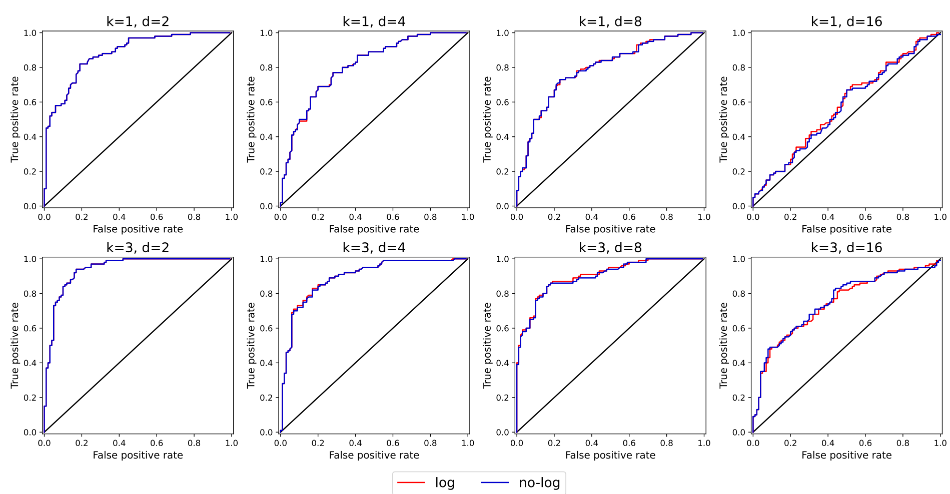
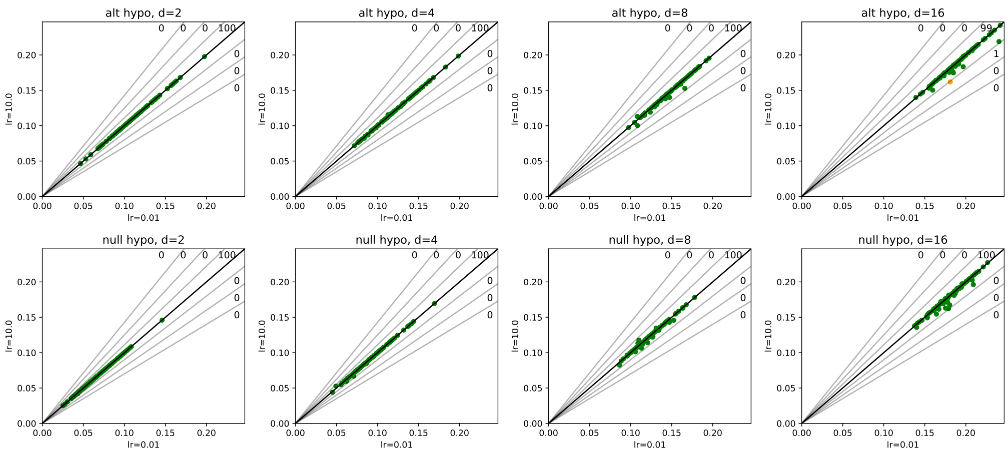
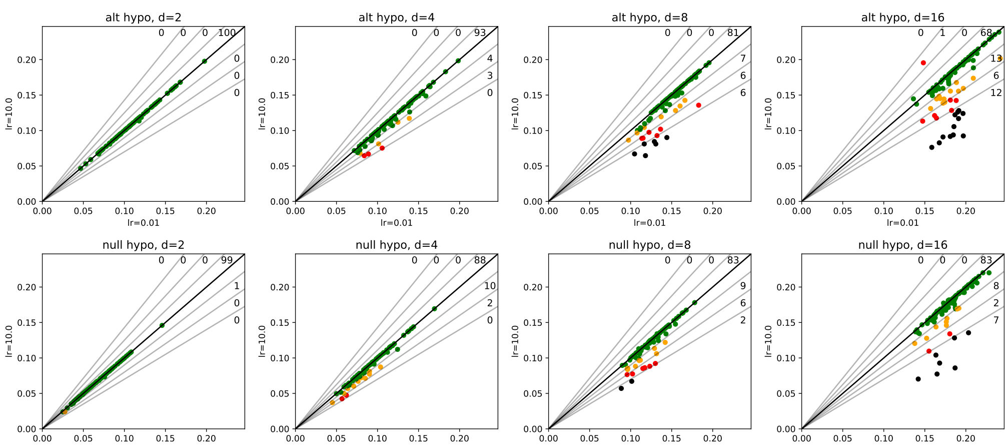


Figure 4: ROC curves from the "log" and "no-log" problems, for the same data as in Figure 3. They are very similar except for the largest k, d pair, where the "no-log" ROC curve is slightly better. Inspecting Figure 3b, this is likely due to the fact that the null MMD values here are relatively smaller (poorer optimization).

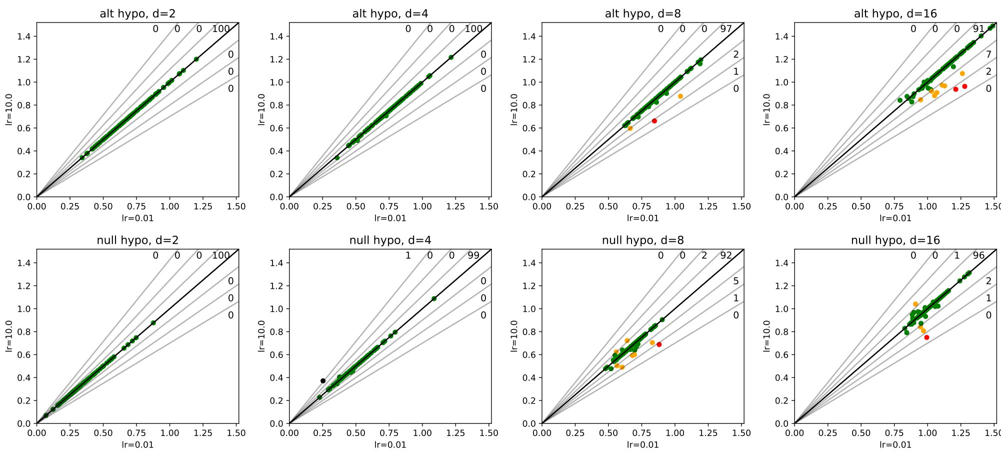


(a) $k = 1$, log

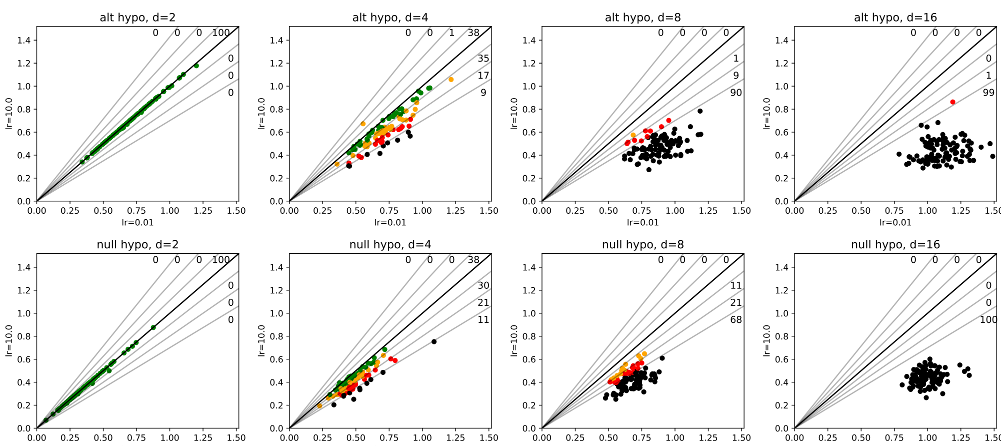


(b) $k = 1$, no-log

Figure 5: MMD values corresponding to learning rates 0.01 (x-axis) and 10 (y-axis), for the same data as in Figure 3. There is little to no difference for the "log" problem (points near the diagonal), and a much bigger difference for the "no-log" problem, where in most cases the larger learning rate is worse (points below the diagonal).

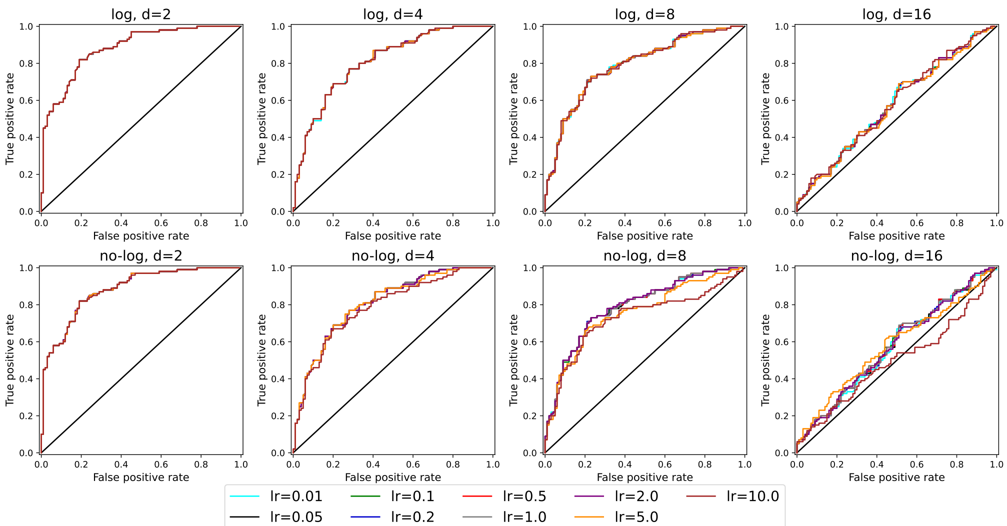


(a) $k = 3$, log

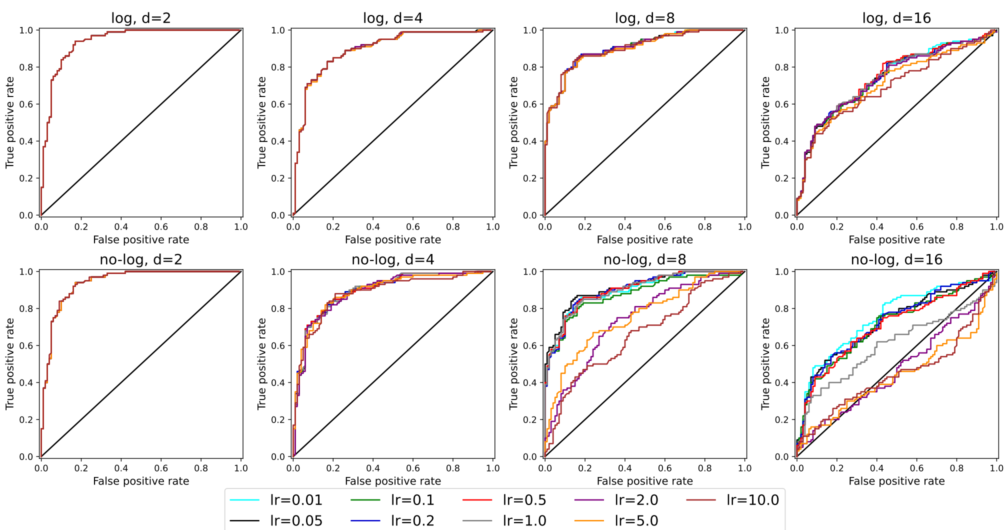


(b) $k = 3$, no-log

Figure 6: MMD values corresponding to learning rates 0.01 (x-axis) and 10 (y-axis), for the same data as in Figure 3. There is little to moderate for the "log" problem (points around the diagonal), and a much bigger difference for the "no-log" problem, where in nearly all cases the larger learning rate is clearly much worse (points below the diagonal).

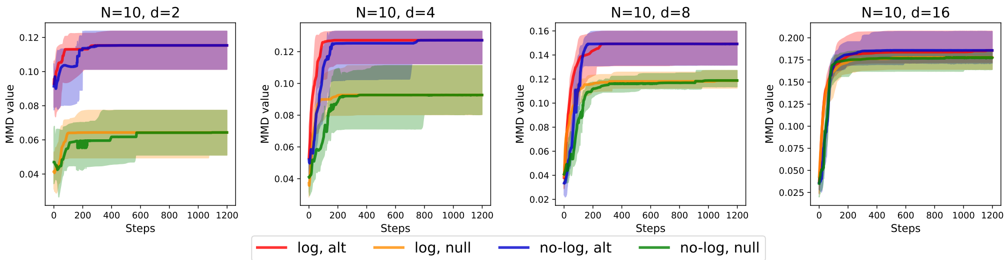


(a) $k = 1$

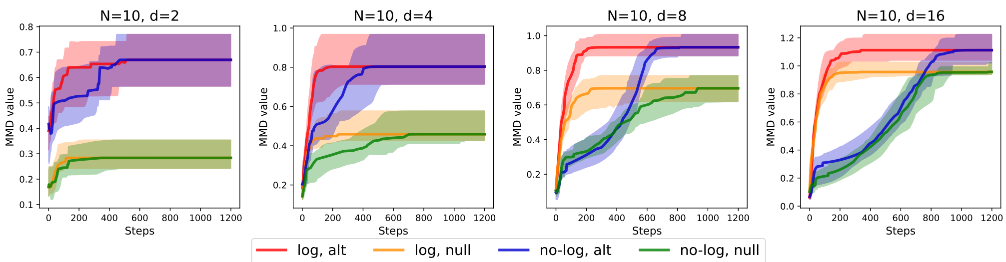


(b) $k = 3$

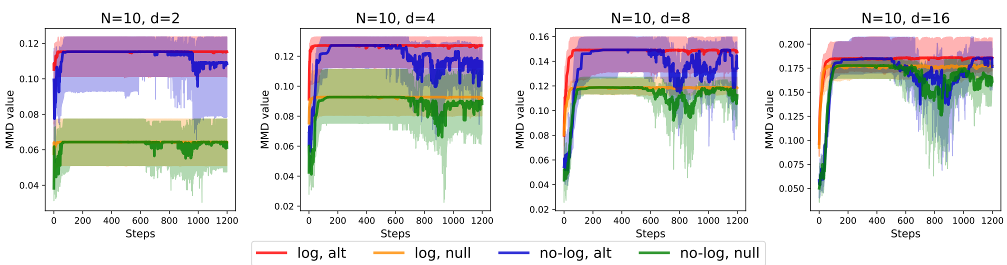
Figure 7: MMD values corresponding to learning rates from 0.01 to 10, for the same data as in Figure 3. In all cases, the "log" problem results in much more stable ROC curves with respect to the choice of learning rate, especially for larger k, d.



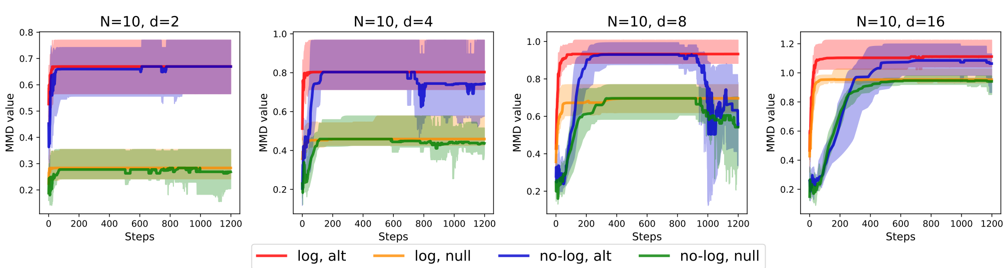
(a) $k = 1$, $\mathrm{lr} = 0.01$



(b) $k = 3$, $\mathrm{lr} = 0.01$

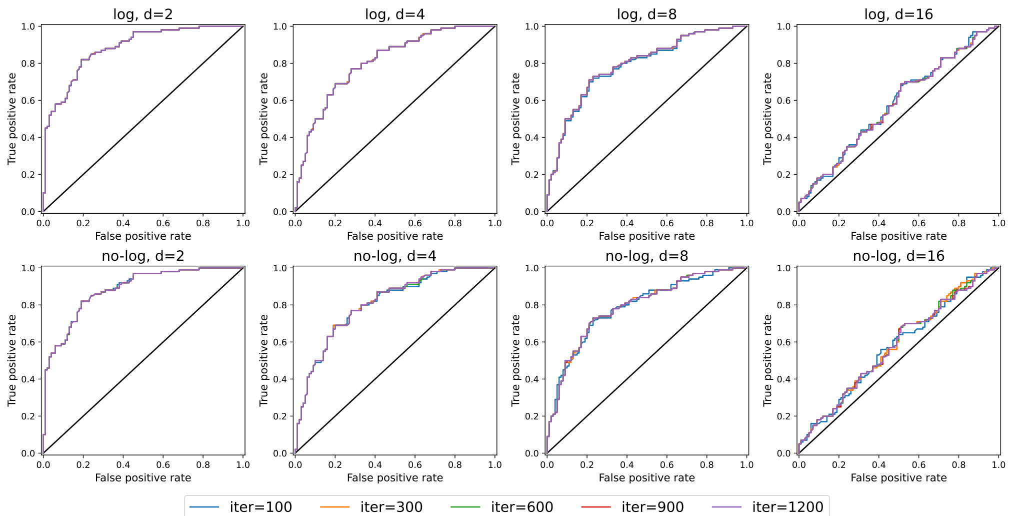


(c) $k = 1$, $\mathrm{lr} = 0.5$

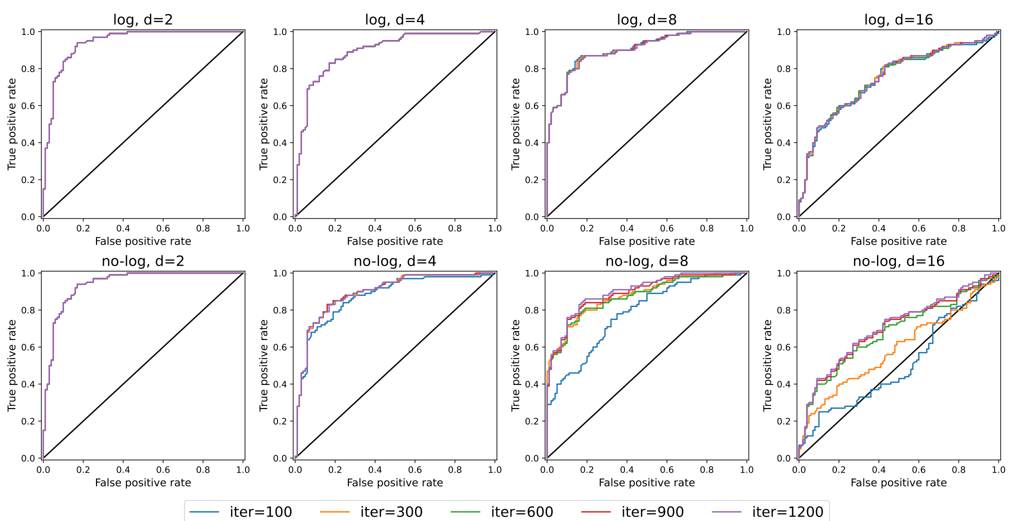


(d) $k = 3$, $\mathrm{lr} = 0.5$

Figure 8: MMD value as a function of iteration for the "log" and "no-log" problems. This is done over 20 repetitions (draws of samples from P, Q); solid lines represent the median and shaded areas the interquartile range (computed over repetitions), at each iteration. For smaller k, d, the "log" and "no-log" convergence speeds are fairly similar, but for larger k, d, the "no-log" convergence speed is clearly slower. For the larger learning rate, we can also see that in many instances the "no-log" iterations fail to converge, and bounce up and down in MMD value, without staying at a local optimum.



(a) $k = 1$



(b) $k = 3$

Figure 9: ROC curves for the "log" and "no-log" problems, at different iteration numbers, over the same data in Figure 8. The "log" curves are quite robust to the choice of the number of iterations, whereas the "no-log" ones are not, especially for larger k, d. In other words, the "log" problem typically shows faster convergence (number of iterations required to approach a local optimum), especially for larger k, d.

J Sensitivity analysis: number of neurons

In this section, we now turn to investigate the role of the number of neurons N. We stick with the "log" problem (9) as in the experiments in the main text (with the previous section providing evidence that the log transform leads to greater robustness in many regards). The setup is essentially the same as that used in the last section, except that we vary the number of neurons from $N = 1$ to $N = 20$. The summary of our findings is as follows.

- Using a larger number of neurons often improves the achieved MMD value, especially for larger k, d; however, going past $N = 10$ does not seem to make a big difference in our experiments.
- Using a larger number of neurons generally results in a better ROC curve, especially for larger k, d; again, going past $N = 10$ does not seem to make a big difference in our experiments.
- Using a larger number of neurons can result in faster convergence, though not dramatically.
- Using a larger number of neurons is typically more robust to the choice of learning rate, both in terms of the MMD values and ROC curves obtained, especially for larger k, d.

Figure 10–14 display the results from our sensitivity analyses. As before, each figure caption presents the salient points about the setup and takeaways, and we remark that results for $k = 4$ (not included for brevity) show qualitatively similar but even more extreme behavior.

K Comparison to polynomial kernel MMD

Figure 15 compares the RKS tests to the polynomial kernel MMD tests using the data as in Figure 2 in the main paper. We have chosen linear, quadratic, and cubic degree polynomials by (rough) analogy to the use of $k = 1, 2, 3$ in RKS. The colors of the RKS test of degree k and kernel MMD with a polynomial kernel of degree k are chosen to match. However, we can see that these two frameworks produce tests with generally quite different behaviors: for example, compare the behavior of RKS with $k = 1$ to kernel MMD with linear kernel, in the "t-coord" setting.

Altogether, the behavior of kernel MMD with polynomial kernel is as expected: it performs quite well when $P \neq Q$ differ according to the alternative (moment difference) that is anticipated by the test, and is otherwise essentially powerless.

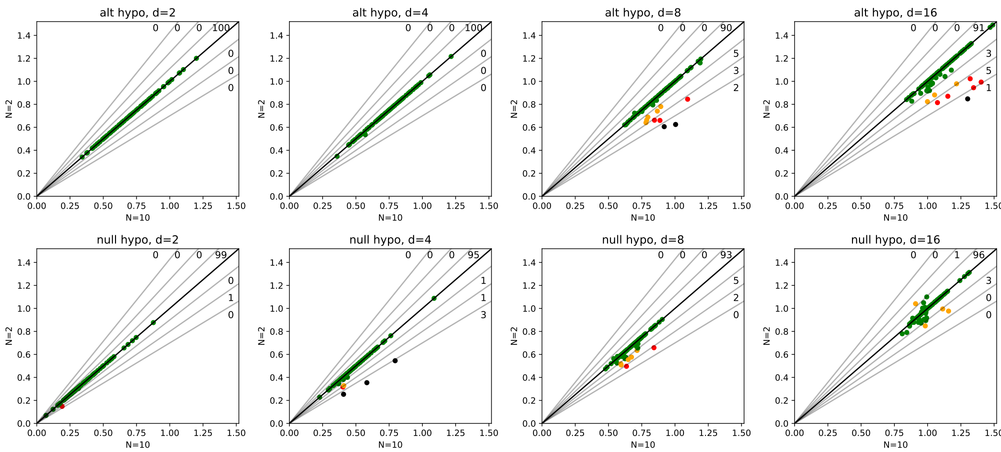


Figure 10: MMD values when $N = 10$ (x-axis) and $N = 2$ (y-axis). Each point represents a set of samples drawn from P, Q, and the result of running $T = 1200$ iterations with learning rate 0.5, when $k = 3$. Points below the diagonal mean that $N = 10$ results in a larger MMD value, which we see is especially prominent for larger k, d, and more prominent under the null.

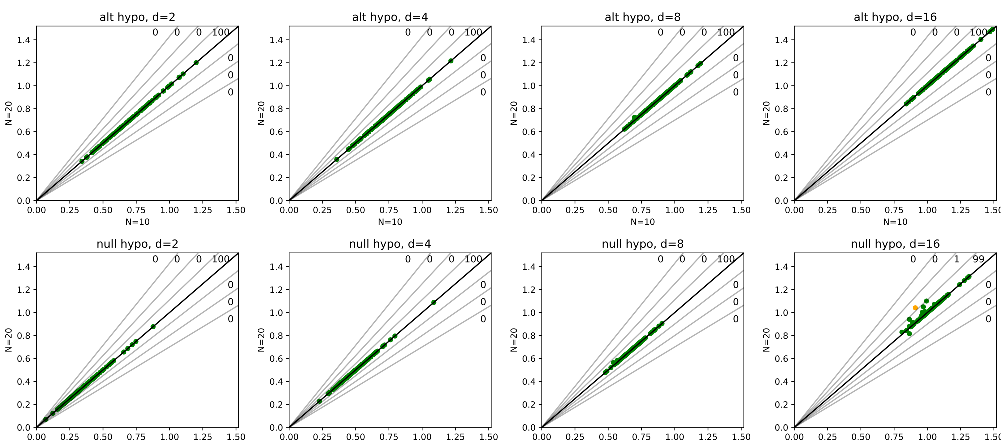


Figure 11: MMD values when $N = 10$ (x-axis) and $N = 20$ (y-axis). The setup is as above. Here we see no clear improvement in moving from $N = 10$ to $N = 20$ neurons.

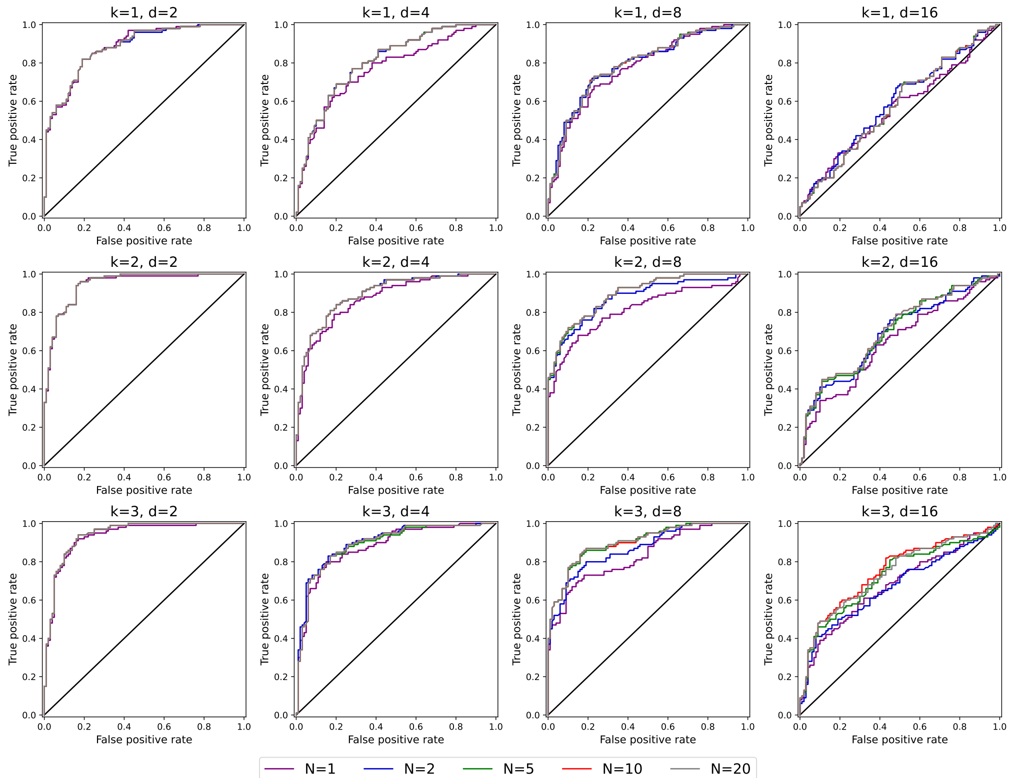
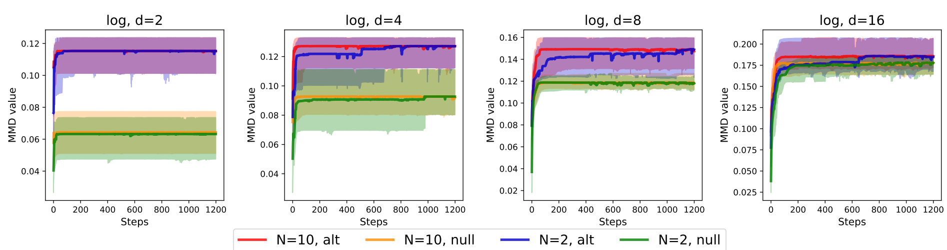
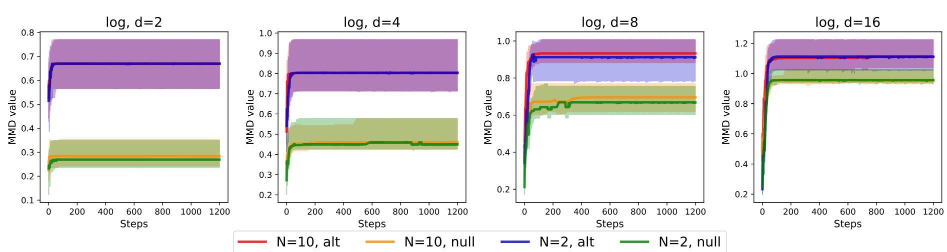


Figure 12: ROC curves as the number of neurons varies from $N = 1$ to $N = 20$, for the same data as in Figure 10. The ROC curves look generally better for a larger number of neurons, especially for larger k, d.



(a) $k = 1$, lr $= 0.5$



(b) $k = 3$, lr $= 0.5$

Figure 13: MMD value as a function of iteration for $N = 10$ and $N = 2$ neurons. This is carried out over 20 repetitions (draws of samples from P, Q); solid lines represent the median and shaded areas the interquartile range (computed over repetitions), at each iteration. The convergence speed is sometimes slightly faster with $N = 10$ neurons but the differences are not large.

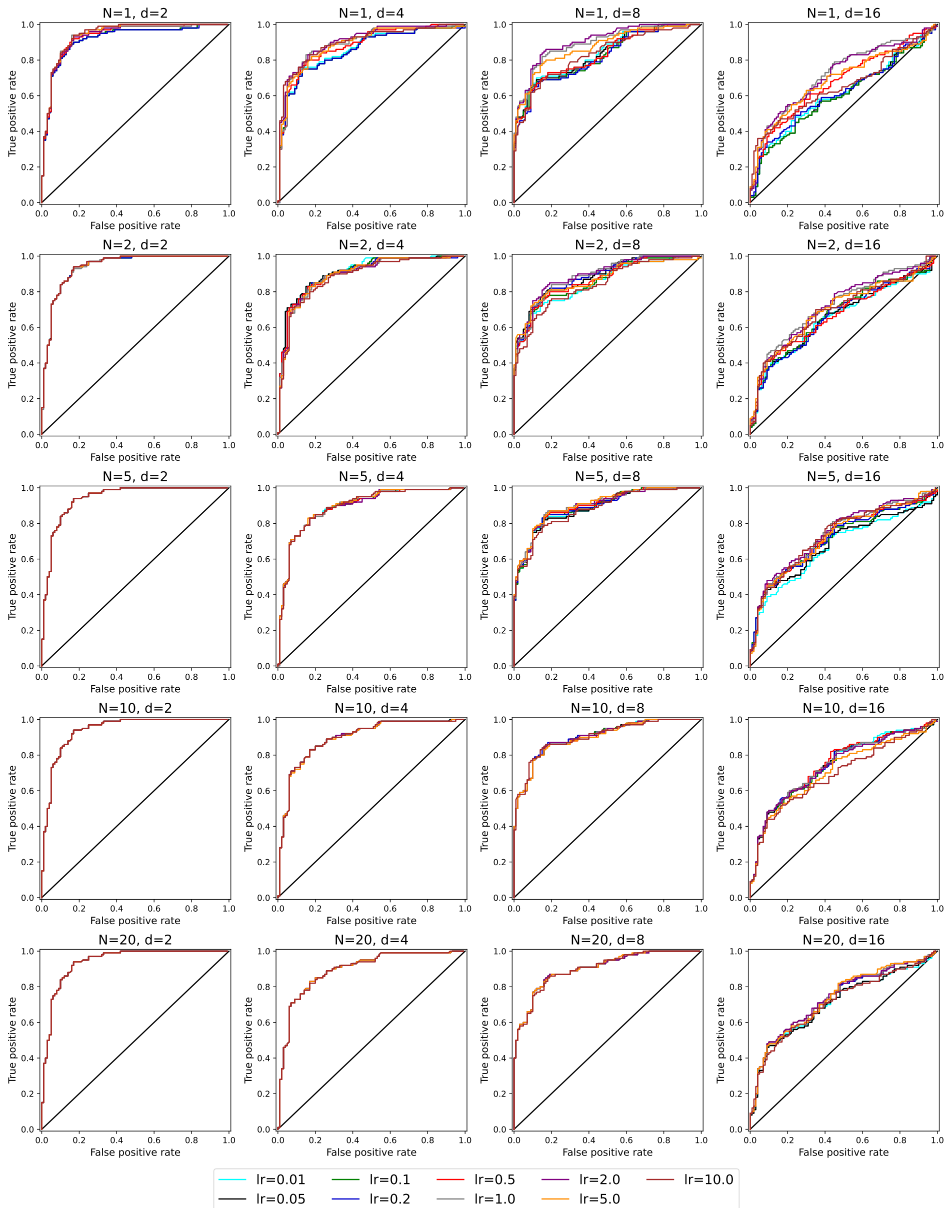


Figure 14: ROC curves as the number of neurons varies from $N = 1$ to $N = 20$, and for varying learning rates, for the same data as in Figure 10. A larger number of neurons is typically more robust to the choice of learning rate, especially for larger d.

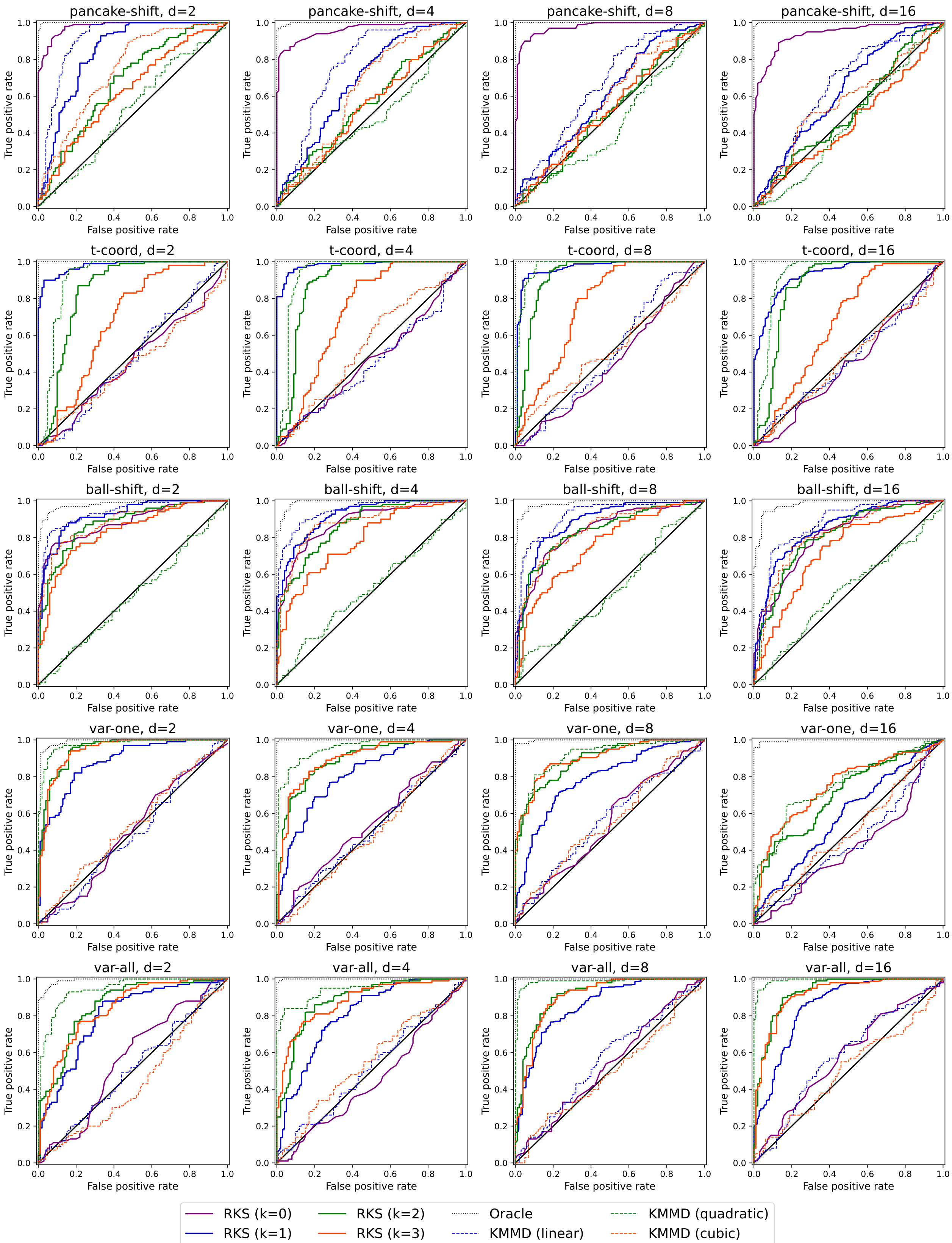


Figure 15: ROC curves for the same data as in Figure 2, but now comparing RKS to kernel MMD (KMMD) with a polynomial kernel of linear, quadratic, and cubic degrees. The ROC curve from the likelihood ratio test is also shown, indicating the best possible performance for each setting (it depends on oracle knowledge of P, Q).