Permutation invariant functions: statistical tests, density estimation, and computationally efficient embedding

Anonymous authors
Paper under double-blind review

Abstract

Permutation invariance is among the most common symmetry that can be exploited to simplify complex problems in machine learning (ML). There has been a tremendous surge of research activities in building permutation invariant ML architectures. However, less attention is given to: (1) how to statistically test for permutation invariance of coordinates in a random vector where the dimension is allowed to grow with the sample size; (2) how to leverage permutation invariance in estimation problems and how does it help reduce dimensions. In this paper, we take a step back and examine these questions in several fundamental problems: (i) testing the assumption of permutation invariance of multivariate distributions; (ii) estimating permutation invariant densities; (iii) analyzing the metric entropy of permutation invariant function classes and compare them with their counterparts without imposing permutation invariance; (iv) deriving an embedding of permutation invariant reproducing kernel Hilbert spaces for efficient computation. In particular, our methods for (i) and (iv) are based on a sorting trick and (ii) is based on an averaging trick. These tricks substantially simplify the exploitation of permutation invariance.

1 Introduction

Many applications can benefit from exploiting a known symmetry in the data. One of the most basic symmetries is permutation invariance, where the function outputs are invariant to the order of the inputs. Permutation invariance plays a crucial role in machine learning applications such as set anomaly detection, text concept set retrieval, and point cloud classification Zaheer et al. (2017). Building permutation invariant machine learning architectures and assessing their computational performance has been a popular topic in the field. In Qi et al. (2017), a point cloud 3D classification model was proposed, where the permutation invariance was built–in using max–pooling. In Zaheer et al. (2017), the authors prove that any permutation invariant function can be written in a certain general form that can be implemented using standard deep learning frameworks, by sum-aggregating the output of the first network and passing into the second network. The general results of Zaheer et al. (2017) provide a basis for other more specialized models such as the attention mechanism for set input Lee et al. (2019). It is noted in Cohen-Karlik et al. (2020) that recurrent neural network has a structure that is naturally suitable as a permutation invariant model using the hidden state as the aggregator, and can be implemented more efficiently than the general form in Zaheer et al. (2017) in some cases. More recently, Tang & Ha (2021) demonstrates that a permutation invariant reinforcement learning agent can learn more robust policies that generalize better to unseen situations. Viewing neural network inputs and outputs as random variables, Bloem-Reddy et al. (2020) studies the structure of neural networks that are useful for modeling data that are invariant, and show a connection between functional and probabilistic symmetry. The representation given by Bloem-Reddy et al. (2020) nests examples from the recent literature, e.g.,Zaheer et al. (2017), as special cases.

The discussion of Bloem-Reddy et al. (2020) raises the fundamental concept, *exchangeability*, which is related to permutation invariance in probability distributions. In particular, a sequence of random variables $X_1, X_2, X_3, \cdots$ is called exchangeable if for any permutation σ, the permuted sequence $X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \cdots$ has the same joint distribution as the original sequence. Using this terminology, a CDF $F(t)$ associated with a sequence of d real–valued random variables $(X_1, ..., X_d) \in \mathbb{R}^d$ is permutation invariant if such a sequence of random variables is exchangeable. Suppose a sample of n i.i.d. random vectors $t_1, t_2, \cdots, t_n$ are drawn from an arbitrary $d-$variate distribution $F(t)$, where $t_i = (t_i^1, t_i^2, \cdots, t_i^d) \in \mathbb{R}^d$ for every $i = 1, ..., n$. Note that $t_1, t_2, \cdots, t_n$ is exchangeable, but each $t_i \sim F$ is not permutation invariant if F is not permutation invariant.

The example above raises the importance of distinguishing *permuting observations* (which makes up a large portion of the literature on "exchangeability") from *permuting the coordinates of a random vector*. There is a vast literature on the former but a relatively scarce literature on the latter (as pointed out by (Kalina & Janáček, 2023, page 3143)). Our paper focuses on the latter. The former has been discussed in the contexts of conformal prediction, testing independence, and testing the equality of two distributions (for example, Kuchibhotla (2020)). There is a separate literature related to permutation tests based on U-statistics such as Chapters 12-13 of Van der Vaart (2000). Permutation test statistics are U-statistics which permute the observations that satisfy i.i.d. or weak dependence conditions. In contrast, our problems permute on the dimension and make *no* assumptions about the dependence among the coordinates of a random vector.

This fact is crucial for numerous applications in health sciences, finance, and climatology where features instead of observations are permuted. For example, one application tests whether the red and white blood cell counts and hemoglobin concentration are permutable in athletes, and measurements were taken from a sample of athletes (Kalina & Janáček (2023)). A researcher may not want to impose any condition on the dependence between different types of measurements. Applications like this one have motivated statisticians to develop tests for the assumption of permutation invariance of coordinates in a random vector; see "Related work" in Section 2.

In this paper, we propose a statistical procedure that tests *directly* whether the coordinates of a random vector from an unknown multivariate distribution are permutable. Specifically, our test statistics take the form $T := \sup_{t \in [0,1]^d} \sqrt{n} \left| \widetilde{F}_n(t) - F_n(t) \right|$, where $F_n(t)$ is the empirical CDF at t, $\widetilde{F}_n(t) = F_n(\operatorname{sort} t)$, and n is the size of the random sample. We approximate the quantile of T with the multiplier bootstrap method. In contrast to the existing procedures for testing permutation invariance in multivariate distributions, our test allows d to grow with the sample size n.

Suppose that our test cannot reject the null hypothesis of permutation invariance. Then one may be interested in estimating the underlying density function by exploiting the potential symmetry. We propose a kernel density estimator (KDE) that averages over a carefully constructed subset of permutations. When the true density is indeed permutation invariant, the averaged KDE yields the same bias as the standard KDE but reduces the variance of the standard KDE by a factor of order $(b^{-\bar{d}}) \wedge \bar{d}$, where $0 < b < 1$ depends on the separation of entries of the point the density is at and $\bar{d}$ is the number of unique entries in that point.

Fundamentally, a class of multivariate functions with permutation invariance has a smaller "size" than without imposing permutation invariance, and consequently, a smaller Radamacher complexity. A measure of "size" is the covering number. As a third contribution, we analyze the covering numbers of two permutation invariant function classes and compare them with their counterparts where permutation invariance is not imposed. We show that the *logarithm* of the covering number for the permutation invariant Hölder class with a boundary condition is reduced by a factor of $d!$. Similarly, for the permutation invariant ellipsoid class, the upper and lower bounds on the *logarithm* of the covering number reduce those of the couterpart without imposing permutation invariance by a factor of $d!$.

Lastly, we study the interpolation and fitting of data points generated by some permutation invariant functions in a reproducing kernel Hilbert space (RKHS), $\mathcal{H}$. Given the positive semidefinite kernel $\mathcal{K}$ associated with $\mathcal{H}$, we propose computing a new kernel $\mathcal{K}^{\text{sort}} := \mathcal{K}(\operatorname{sort} ., \operatorname{sort} .)$. We bound the error from approximating a permutation invariant function in $\mathcal{H}$ with a function constructed based on $\mathcal{K}^{\text{sort}}$. This result provides a computationally efficient embedding. Compared with previous proposals such as Bietti et al. (2021); Klus et al. (2021); Tahmasebi & Jegelka (2023) where computing the permutation invariant kernel costs $d!$, our sorted kernel only takes $O(d \log d)$ to compute. At the expense of the computational efficiency, a loss of accuracy can occur in some situations but is insignificant in others. We examine such a trade–off in numerical experiments.

Notation. For two functions $f(n,\gamma)$ and $g(n,\gamma)$, let us write $f(n,\gamma) \gtrsim g(n,\gamma)$ if $f(n,\gamma) \geq cg(n,\gamma)$ for a universal constant $c \in (0, \infty)$; similarly, we write $f(n,\gamma) \lesssim g(n,\gamma)$ if $f(n,\gamma) \leq c^{'} g(n,\gamma)$ for a universal constant $c^{'} \in (0, \infty)$; and $f(n,\gamma) \asymp g(n,\gamma)$ if $f(n,\gamma) \gtrsim g(n,\gamma)$ and $f(n,\gamma) \lesssim g(n,\gamma)$.

Definition 1.1 A permutation invariant function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that $f(\sigma(t)) = f(t)$ for any permutation σ and $t \in \mathbb{R}^d$. We write S_d to denote the set of all permutations.

2 Testing permutation invariance with sorting

Let us consider a random sample consisting of i.i.d. entries $\{t_i \in [0,1]^d\}_{i=1}^n$ from an unknown distribution F over $[0,1]^d$, where potentially the dimension $d \to \infty$ as $n \to \infty$. In this section, we are interested in testing the hypothesis:

$$\begin{cases} H_0 : F \text{ is permutation invariant} \\ H_1 : F \text{ is not permutation invariant} \end{cases}.$$

Our test leverages the following proposition.

Proposition 2.1 *A function f on $[0,1]^d$ is permutation invariant if and only if $f(\operatorname{sort} t) = f(t)$ for all $t \in [0,1]^d$.*

Proof: Suppose f is permutation invariant. Then, for any $t \in [0,1]^d$, there exists $\sigma^* \in S_d$ such that $\operatorname{sort} t = \sigma^*(t)$. Consequently, $f(\operatorname{sort} t) = f(\sigma^*(t)) = f(t)$. Suppose that $f(\operatorname{sort} t) = f(t)$ for all $t \in [0,1]^d$. Then, for any $\sigma \in S_d$, we have $f(\sigma(t)) = f(\operatorname{sort} \sigma(t)) = f(\operatorname{sort} t) = f(t)$. $\square$

The multiplier bootstrap test with a sorting trick. We define the empirical CDF

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n}[t_i \leq t]$$

and the *sorted* empirical CDF

$$\widetilde{F}_n(t) := F_n(\operatorname{sort} t) = \frac{1}{n}\sum_{i=1}^{n}[t_i \leq \operatorname{sort} t].$$

Given $t = (t^1, \cdots, t^d) \in [0,1]^d$, we define $\operatorname{sort} t := (t^{\pi(1)}, \cdots, t^{\pi(d)})$ for some permutation $\pi \in S_d$ such that $0 \leq t^{\pi(1)} \leq t^{\pi(2)} \leq \cdots \leq t^{\pi(d)} \leq 1$.[1]

We propose the following statistics

$$T := \sup_{t \in [0,1]^d} \sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right|$$

and the multiplier bootstrap version

$$W := \sup_{t \in [0,1]^d} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\left([t_i \leq \operatorname{sort} t] - [t_i \leq t]\right) e_i\right|, \quad e_i \sim_{iid} \mathcal{N}(0,1)$$

along with the corresponding bootstrap critical value

$$c_W(\alpha) := \inf\left\{t \in \mathbb{R} \; : \; \mathbb{P}_e[W \leq t] \geq 1 - \alpha\right\}.$$

We establish the following theoretical guarantee for our test.

Theorem 2.2 *Suppose that $d = o\left(n^{1/7}\right)$ and the CDF F is continuous. Under H_0, there exists some universal constants $c, C \in (0,\infty)$, such that*

$$\sup_{\alpha \in (0,1)}\left|\mathbb{P}\left[\sup_{t \in [0,1]^d} \sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right| > c_W(\alpha)\right] - \alpha\right| < Cn^{-c} \to 0 \text{ as } n \to \infty.$$

[1]In this paper, $t = (t^1, \cdots, t^d)$ denotes a point (a $d-$dimensional vector) a function is evaluated at, whereas $\{t_i \in [0,1]^d\}_{i=1}^n$ denotes a random sample drawn from some $d-$variate probability distribution.

Remark. Note that we make *no* assumptions about the dependence among the coordinates of a random vector $t_i = (t_i^1, ...t_i^d)$ for any $i = 1, ..., n$, even though the sample $\{t_i \in [0,1]^d\}_{i=1}^n$ consists of i.i.d. observations.

Sketch of the proof. Suppose that we have a list of points $\{v_j\}_{j=1,\cdots,N}$ in $[0,1]^d$ which is sufficiently large and well chosen. We should be able to approximate $\sup_{t\in[0,1]^d} \sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right|$ by the maximum of the coordinates of

$$\left(\sqrt{n}\left|\widetilde{F}_n(v_1) - F_n(v_1)\right|, \cdots, \sqrt{n}\left|\widetilde{F}_n(v_N) - F_n(v_N)\right|\right).$$

The above can be expressed using a sum of independent random vectors. From there, we apply the result of Chernozhukov et al. (2013). So, the key to the proof is to construct a desired list of points $\{v_j\}$. We let $\{v_j\}$ be the points on a $n^m \times \cdots \times n^m$ grid on $[0,1]^d$ for some $m \geq 4$ and argue: when the grid is fine enough, the probability that the supremum is reached at one of the n^{md} grid point approaches one sufficiently quickly.

The full proof can be found in Section A.1 of the appendix.

Implementation. In practice, given the data $\{t_i\}_{i=1}^n$, we estimate the supremum

$$\sup_{t\in[0,1]^d} \sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right|.$$

Similarly, computing the supremum W for N_W draws of $\{e_i\}$, one can numerically estimate $c_W(\alpha)$. The test in Theorem 2.2 rejects H_0 at $\alpha \in (0,1)$ significance level (e.g., $\alpha = 0.05$ or 0.01) if $\sup_{t\in[0,1]^d} \sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right| > c_W(\alpha)$. There are a number of ways the supremum of

$$Z_n(t; \{e_i\}_{i=1}^n) := \left|\sum_{i=1}^n \left([t_i \leq \operatorname{sort} t] - [t_i \leq t]\right) e_i\right|$$

(i.e., W) can be estimated. The problem of finding the supremum of $\left|\widetilde{F}_n(t) - F_n(t)\right|$ is the special case where $e_i = 1$ for all $i = 1, \cdots, n$, and we write $Z_n(t) := Z_n(t; \{e_i = 1\}_{i=1}^n)$ for convenience.

The supremum of $Z_n(t; \{e_i\}_{i=1}^n)$ can be reached at some unsorted point t^* of the form

$$t^* = \left(t_{i_1}^{a_1}, \cdots, t_{i_d}^{a_d}\right)$$

for some $a_1, \cdots, a_d \in \{1, \cdots, d\}$ and $i_1, \cdots, i_d \in \{1, \cdots, n\}$. This can be seen from that given $t, t' \in [0,1]^d$, $Z_n(t; \{e_i\}_{i=1}^n)$ and $Z_n(t'; \{e_i\}_{i=1}^n)$ can differ from each other by no more than $\frac{1}{\sqrt{n}}$ multiple of the sum of the number of times each coordinate of t' needs to *cross* any of the nd numbers $\{t_i^a\}$ to get to t:

$$|Z_n(t; \{e_i\}_{i=1}^n) - Z_n(t'; \{e_i\}_{i=1}^n)| \leq \frac{1}{\sqrt{n}} \sum_{b=1}^d \left|\{t_i^a\}_{a=1,\cdots,d; i=1,\cdots,n} \cap [t^b, t'^b]\right|. \tag{1}$$

Therefore, the upper bound estimate for the number of points needed to be searched to find the supremum is $(nd)^d$. When n and d are large, this upper bound is unfavorable.

We propose a more practical solution by first defining a smoothened version of the CDF, given $\varepsilon > 0$:

$$F_{n,h}(t) := \frac{1}{n}\sum_{i=1}^n \frac{1}{2^d}\prod_{a=1}^d \left(\tanh\left(\frac{t^a - t_i^a}{h}\right) + 1\right)$$

and $\widetilde{F}_{n,h}(t) := F_{n,h}(\operatorname{sort} t)$. The parameter h controls smoothness, with $F_{n,h}$ and $\widetilde{F}_{n,h}$ converges in L^1 to their empirical counter–part. The added smoothness provides the needed gradient for a standard maximization algorithm to approach the supremum given a randomly chosen starting point. For our implementation, we use COBYLA, $h = 0.001$, and $n/2$ random starting points, which work well from our observation.

Remark. Via experimentation, we observed that W (and T) can be approached by searching for the maximum of $Z_n(t;\{e_i\}_{i=1}^n)$ over a set $\{v_j\}_{j=1,\cdots,(nd^2)^q}$ of $(nd^2)^q$ test points randomly drawn from F. Here $q \in (1,\infty)$ is an arbitrary constant and $(nd^2)^q$ should be interpreted as $\lceil (nd^2)^q \rceil$, or any other similar convention, if q is not an integer. In other words, we expect $\max_{j=1,\cdots,(nd^2)^q} Z_n(v_j)$ and $\max_{j=1,\cdots,(nd^2)^q} Z_n(v_j;\{e_i\}_{i=1}^n)$ to converge to T and W *sufficiently quickly* (in the sense that is needed by Chernozhukov et al. (2013), see (9) for the technical detail) as $nd^2 \to \infty$. The rate of convergence depends on the choice of q, and it is left to the practitioner to decide q that balances the computational resource and performance. Figure 1 in Appendix A.9 exhibits the convergence to T in the case of $d = 2$ with $q = 1.5$, and suggests that the number of the points needed in the proof of Theorem 2.2 may grow much slower than n^{md}, in particular, possibly be only polynomial in d. This evidence prompts the following open question.

Question. Will the statement of Theorem 2.2 holds for $d = o\left(e^{n^{1/7}}\right)$?

Related work. The first bivariate symmetry test was proposed by Hollander (1971), motivated by the question of whether a medical treatment improves patient conditions. Also see Lyu & Belalia (2023); Yanagimoto & Sibuya (1976) for other proposals. More recently, tests of permutation invariance for distributions with more than two dimensions have been proposed in Bahraoui & Quessy (2022); Harder & Stadtmüller (2017); Kalina & Janáček (2023). The procedure proposed in Kalina & Janáček (2023) tests the null hypothesis of *pairwise* symmetry instead of permutation invariance per se. As acknowledged in Kalina & Janáček (2023), pairwise symmetry is a weaker condition than permutation invariance. Related is a test of a weaker condition of permutation invariance for multivariate copulas, which has been studied in Bahraoui & Quessy (2022); Harder & Stadtmüller (2017). In contrast to this literature, our procedure tests *directly* whether the coordinates of a random vector from an unknown multivariate probability distribution are permutable.

3 Estimating permutation invariant densities with averaging

In Section 2, we have proposed a statistical procedure for testing permutation invariance of multivariate distributions. In this section, we focus on permutation invariant density functions and show a way to exploit permutation variance in the (local) kernel density estimation method (see, e.g., Tsybakov (2009)).

The averaging trick. Given $t = (t^1, \cdots, t^d) \in \mathbb{R}^d$, the *standard* kernel density estimator is given by

$$\hat{f}(t) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{t - t_i}{h}\right).$$

To exploit the permutation invariance structure, we propose the *averaged* kernel density estimator

$$\tilde{f}(t) = \frac{1}{\bar{d}} \sum_{\sigma \in S_{\bar{d}}^*} \hat{f}(\sigma(t)),$$

where $\bar{d}$ and $S_{\bar{d}}^*$ are defined as follows.

- Let $t \in \mathbb{R}^d$ be a vector with $\bar{d}$ distinct entries. The set $S_{\bar{d}}^*$ can be *any* set consisting of $\bar{d}$ permutations such that any $\sigma, \pi \in S_{\bar{d}}^*$ take different values on exactly $\bar{d}$ positions. When $\bar{d} < d$, a set satisfying the aforementioned property is not unique.

For example, take $t = \{1, 2, 3\}$. Then $\bar{d} = 3$ and $S_{\bar{d}}^* = \{\{1, 2, 3\}, \{3, 1, 2\}, \{2, 3, 1\}\}$. As another example, take $t = \{1, 1, 2\}$. Then $\bar{d} = 2$ and the choice of $S_{\bar{d}}^*$ can be any of the three sets: $\{\{1, 1, 2\}, \{2, 1, 1\}\}$, $\{\{1, 1, 2\}, \{1, 2, 1\}\}$, $\{\{2, 1, 1\}, \{1, 2, 1\}\}$. Clearly, when all entries of t take the same value, $\bar{d} = 0$ and $\hat{f}(t) = \tilde{f}(t)$.

Our theoretical guarantees are based on the following assumption.

Assumption 3.1 The random sample consisting of i.i.d. entries $\{t_i \in [0,1]^d\}_{i=1}^n$ is from a pdf f, which is permutation invariant and twice differentiable with bounded derivatives. The non-negative kernel K satisfies: (a) $\int_{-\infty}^{\infty} K(v)dv = 1$; (b) $K(v) = K(-v)$ for all v; (c) $\int_{-\infty}^{\infty} vv^T K(v)dv < \infty$; (d)$\int_{-\infty}^{\infty} K(v)^2 dv < \infty$.

In what follows, we compare the bias and variance of $\tilde{f}$ with those of $\hat{f}$. When writing *higher order terms*, we mean that these terms have a smaller order than the leading term(s).

Lemma 3.2 *Let Assumption 3.1(a-c) hold. Suppose $h \to 0$. Then,*

$$\mathbb{E}[\hat{f}(t)] - f(t) = \mathbb{E}[\tilde{f}(t)] - f(t) = \frac{h^2}{2} tr\left(\frac{\partial^2 f(t)}{\partial t \partial t^T} \int vv^T K(v)dv\right) + \textit{higher order terms.}$$

This result shows that the biases of $\tilde{f}$ and $\hat{f}$ have the same leading term.

Lemma 3.3 *Let Assumption 3.1 hold. Suppose $n \to \infty$ and $h \to 0$ while $nh^d \to \infty$.*

(i) Then, the variances are computed as

$$\mathbb{V}(\hat{f}(t)) = \frac{f(t)}{nh^d} \int_{-\infty}^{\infty} K(v)^2 dv + \textit{higher order terms} \tag{2}$$

and

$$\mathbb{V}(\tilde{f}(t)) = \frac{f(t)}{(\bar{d})^2 nh^d} \sum_{\pi,\sigma \in S_{\bar{d}}^* \ s.t.\ \pi \neq \sigma} K * K\Big(\frac{\sigma(t) - \pi(t)}{h}\Big) + \frac{1}{\bar{d}} \frac{f(t)}{nh^d} \int_{-\infty}^{\infty} K(v)^2 dv + \ \textit{higher order terms} \tag{3}$$

where $*$ denotes the convolution.

(ii) Further, if we consider the product kernel $K(v) = k(v^1)k(v^2)...k(v^d)$ where $k(\cdot)$ is a non-negative univariate kernel satisfying conditions (a)-(d) in Assumption 3.1 and $k(v)$ decreases in $|v|$,[2] then

$$K * K\Big(\frac{\sigma(t) - \pi(t)}{h}\Big) \leq b^{\bar{d}} \int_{-\infty}^{\infty} K(v)^2 dv \tag{4}$$

where $0 < b < 1$ depends on $\frac{\sigma(t)-\pi(t)}{h}$.

The proofs of Lemma 3.2 and Lemma 3.3 can be found in Section A.2 and Section A.3 of the appendix, respectively.

The reduction in variance and mean squared errors. The result in (2) is the standard one in the literature. From (4) and the first two terms in (3), compared to the standard product kernel density estimator $\hat{f}(t)$, we see a reduction in the variance of the *averaged* kernel density estimator $\tilde{f}(t)$ when $\bar{d} \geq 1$. The larger $\bar{d}$ and $\frac{\sigma(t)-\pi(t)}{h}$ are, the greater the reduction is. Given that the point-wise mean square error $\mathrm{MSE}(t) = \mathbb{V}(t) + (\mathrm{Bias}(t))^2$ and Lemma 3.2, the reduction in the variance implies that $\tilde{f}(t)$ has a smaller $\mathrm{MSE}(t)$ than $\hat{f}(t)$ when $\bar{d} \geq 1$. Consequently, $\tilde{f}$ also has a smaller mean integrated squared error (MISE) than $\hat{f}$.

4 A fundamental perspective

Fundamentally, a class of multivariate functions with permutation invariance has a smaller "size" than without imposing permutation invariance, and consequently, a smaller Radamacher complexity. A measure of "size" is the metric entropy such as covering numbers; see, for example, Chapters 5 and 13 of Wainwright

[2]Examples of kernels satisfying these assumptions include the triangular kernel, the Gaussian kernel, the cosine kernel, the Epanechnikov kernel, the quartic kernel, the triweight kernel, the tricube kernel, the logistic kernel, the sigmoid function and etc.

(2019) for the relationship between Radamacher/Gaussian complexity and metric entropy. In the following, we compare the metric entropy of two permutation invariant function classes with the metric entropy of their counterparts where permutation invariance is not imposed.

Let $p = (p_j)_{j=1}^{d}$ and $P = \sum_{j=1}^{d} p_j$ where p_js are non-negative integers; $x = (x_j)_{j=1}^{d}$ and $x^p = \prod_{j=1}^{d} x_j^{p_j}$. Write $D^p f(x) = \partial^P f / \partial x_1^{p_1} \cdots \partial x_d^{p_d}$.[3]

Definition 4.1 [Hölder classes with a boundary condition] For a non-negative integer γ, let the permutation invariant Hölder class $\mathcal{U}^{perm}$ be the class of functions such that any function $f \in \mathcal{U}^{perm}$ satisfies: (1) f is continuous and permutation invariant on $[0,\, 1]^d$, and all partial derivatives of f exist for all p with $P := \sum_{k=1}^{d} p_k \leq \gamma$; (2) $|D^p f(x)| \leq C$ for all p with $P = k$ $(k = 0, ..., \gamma)$ and $x \in [0,\, 1]^d$ such that $D^p f(0) = 0$ (the boundary condition), where $D^0 f(x) = f(x)$; (3) $\left| D^p f(x) - D^p f(x^{'}) \right| \leq C \left| x - x^{'} \right|_{\infty}$ for all p with $P = \gamma$ and $x,\, x^{'} \in [0,\, 1]^d$. When permutation invariance is not imposed, we denote the Hölder class by $\mathcal{U}$.

Theorem 4.2 *We have*

$$\log N_2(\delta, \mathcal{U}^{perm}) \asymp \log N_\infty(\delta, \mathcal{U}^{perm}) \asymp \frac{1}{d!} \log N_\infty(\delta, \mathcal{U}) \asymp \frac{1}{d!} b_{d,\gamma}^{d} \delta^{\frac{-d}{\gamma+1}}$$

where $b_{d,\gamma}$ is a function of (d, γ) only and independent of δ, $N_2(\delta, \mathcal{U}^{perm})$ denotes the $\delta-$covering number of $\mathcal{U}^{perm}$ with respect to the L^2-norm and $N_\infty(\delta, \mathcal{U}^{perm})$ ($N_\infty(\delta, \mathcal{U})$) denotes the $\delta-$covering number of $\mathcal{U}^{perm}$ (respectively, $\mathcal{U}$) with respect to the sup norm.

The proof of Theorem 4.2 can be found in Section A.4 of the appendix.

Definition 4.3 [Ellipsoid classes] Given a sequence of non–negative real numbers $\{\mu_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$ such that $\sum_{k \in \mathbb{Z}_{\geq 0}^d} \mu_k < \infty$, we define the ellipsoid

$$\mathcal{E} := \left\{ (\beta_k)_{k \in \mathbb{Z}_{\geq 0}^d} \mid \sum_{k \in \mathbb{Z}_{\geq 0}^d} \frac{\beta_k^2}{\mu_k} \leq 1 \right\},$$

and its permutation invariant subset:

$$\mathcal{E}^{perm} := \left\{ (\beta_k)_{k \in \mathbb{Z}_{\geq 0}^d} \in \mathcal{E} \mid \beta_{\mathrm{sort}\, k} = \beta_k \right\}.$$

Consider an RKHS $\mathcal{H}$ of functions over $[0,1]^d$ with Mercer's kernel $\mathcal{K}$, whose associated eigenfunctions $\{\phi_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$ satisfy $\phi_k(\sigma t) = \phi_{\sigma^{-1} k}(t)$ for all $\sigma \in S_d$, and we denote the associated eigenvalues by $\{\mu_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$. By definition, $\{\phi_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$ gives an orthonormal basis for $L^2([0,1]^d, \mathbb{P})$. Since the kernel is continuous on the compact domain $[0,1]^d$, we have the convergence $\sum_{k \in \mathbb{Z}_{\geq 0}^d} \mu_k = \int_{[0,1]^d} \mathcal{K}(t,t) d\mathbb{P}(t) < \infty$. It is well known (see, e.g., Corollary 12.26 Wainwright (2019)) that $\mathcal{H}$ can be identified with the ellipsoid $\mathcal{E}$ where any $f \in \mathcal{H}$ can be written in the form $f = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k$. Now, consider the subspace of permutation invariant functions

$$\mathcal{H}^{perm} := \{ f \in \mathcal{H} \mid f(\mathrm{sort}\, t) = f(t), \ \forall t \} \subset \mathcal{H}.$$

It follows that if $f = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k \in \mathcal{H}^{perm}$, then

$$\sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k(t) = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k(\sigma t) = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_{\sigma k} \phi_{\sigma k}(\sigma t) = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_{\sigma k} \phi_k(t)$$

[3] We use t in most of the results in this paper but use x here to avoid confusion in the notation.

for all t and σ. Hence, $\beta_{\text{sort}\,k} = \beta_k$ and $\mathcal{H}^{perm}$ can be identified with $\mathcal{E}^{perm}$.

Define the norms $\|.\|_{l^2}$ and $\|.\|_{l^2}^{perm}$ on $\mathcal{E}$ and $\mathcal{E}^{perm}$,

$$\|\beta - \beta'\|_{l^2} := \sqrt{\sum_{k\in\mathbb{Z}^d_{\geq 0}} (\beta_k - \beta'_k)^2}, \quad \|\beta - \beta'\|_{l^2}^{perm} := \sqrt{\sum_{k\in\text{sort}\,\mathbb{Z}^d_{\geq 0}} (\beta_k - \beta'_k)^2},$$

for any $\beta, \beta' \in l^2(\mathbb{Z}^d_{\geq 0})$. The following result shows the reduction in metric entropy from imposing permutation invariance.

Theorem 4.4 *There exists $\underline{g}, \bar{g} : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ such that*

$$\underline{g}(\delta) \lesssim \log N(\delta, \mathcal{E}, \|.\|_{l^2}) \lesssim \bar{g}(\delta)$$

and

$$\frac{1}{d!}\underline{g}(\delta) \lesssim \log N(\delta, \mathcal{E}^{perm}, \|.\|_{l^2}^{perm}) \lesssim \frac{1}{d!}\bar{g}(\delta)$$

where $N(\delta, \mathcal{E}, \|.\|_{l^2})$ denotes the $\delta-$covering number of $\mathcal{E}$ with respect to the $\|.\|_{l^2}-$norm and $N(\delta, \mathcal{E}^{perm}, \|.\|_{l^2}^{perm})$ denotes the $\delta-$covering number of $\mathcal{E}^{perm}$ with respect to the $\|.\|_{l^2}^{perm}-$norm.

The proof of Theorem 4.4 can be found in Section A.5 of the appendix.

Like $\mathcal{U}^{perm}$, both the lower and upper bounds on the logarithm of the covering number for $\mathcal{E}^{perm}$ are reduced by a factor of $d!$ when permutation invariance is imposed.

An example. If $\mathcal{H} = W^{s,2}_{per}([0,1]^d)$, the periodic Sobolev space (which is an RKHS when $s > d/2$), it is possible to show that $\bar{g}(\delta) = \underline{g}(\delta) := \left(\frac{1}{\delta}\right)^{d/s}$ in Theorem 4.4. Therefore, we have the sharp result: $\log N(\delta, \mathcal{E}^{perm}, \|, \|_{l^2}^{perm}) \asymp \frac{1}{d!}\log N(\delta, \mathcal{E}, \|, \|_{l^2})$. In Appendix A.6, we derive the reproducing kernel of $W^{s,2}_{per}([0,1]^d)$, i.e. the subspace of the Sobolev space $W^{s,2}([0,1]^d)$ for functions f such that $f(t+k) = f(t)$ for any $k \in \mathbb{Z}^d$.

Related work. To our best knowledge, the most relevant results on metric entropy calculation in the literature of invariant learning are Chen et al. (2023); Sokolic et al. (2017). However, these results concern a different type of symmetry that requires the assumption $\|\sigma(t) - \sigma'(t)\|_2 > 2\delta$ for all the underlying symmetry transformations $\sigma \neq \sigma'$ in Chen et al. (2023); Sokolic et al. (2017). This assumption is neither desirable nor needed for our results; for example, for all $t \in [0,1]^d$ such that all entries are the same, this assumption would be too restrictive (to our setup) as it does not allow $\sigma(t) = t$ for some non-identity permutation σ.

5 Function interpolation with sorting

In this section, we are interested in the interpolation and fitting of data points generated by some permutation invariant functions in an RKHS $\mathcal{H}$. We assume that $\mathcal{H}$ is equipped with the symmetric positive semidefinite kernel $\mathcal{K}$ that satisfies $\mathcal{K}(t,t') = \mathcal{K}(\sigma t, \sigma t')$ for all $\sigma \in S_d$, and is 2ν–continuously differentiable, namely, $\mathcal{K} \in C^{2\nu}([0,1]^d \times [0,1]^d)$.

First, let us review the problem of minimal norm interpolation of the values $y_i = f(t_i), i = 1, \cdots, n$, sampled from an unknown d–variate function $f : [0,1]^d \to \mathbb{R}$ over the design sequence $\{t_i\}_{i=1}^n$. Formally, this problem can be formulated as: $\hat{f} \in \arg\min_{\check{f}\in\mathcal{H}} \|\check{f}\|_{\mathcal{H}}$ subject to $\check{f}(t_i) = y_i,\ i = 1, \cdots, n$. If $f \in \mathcal{H}$, then the solution is always possible by applying the orthogonal projection onto the closed subspace spanned by $\{\mathcal{K}(., t_i)\}_{i=1}^n$, and hence we can write $\hat{f} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\pi}_i \mathcal{K}(., t_i)$ for some $\hat{\pi} \in \mathbb{R}^n$. To learn a permutation invariant function $f \in \mathcal{H}$, a natural approach is to consider the permutation invariant RKHS $\mathcal{H}^{perm}$, generated by the kernel $\mathcal{K}^{perm} := \frac{1}{(d!)^2}\sum_{\sigma,\sigma'\in S_d} \mathcal{K}(\sigma., \sigma'.) = \frac{1}{d!}\sum_{\sigma\in S_d} \mathcal{K}(\sigma., .)$. We refer the readers to Bietti et al. (2021); Klus et al. (2021); Tahmasebi & Jegelka (2023) for previous works related to $\mathcal{H}^{perm}$. For completeness, we summarize the relevant result as follows (see Section A.7 in the appendix for the proof).

Lemma 5.1 *The permutation invariant RKHS $\mathcal{H}^{perm}$ is the subspace of permutation invariant functions of $\mathcal{H}$ with the inner product given by $\langle .,. \rangle_{\mathcal{H}^{perm}} = \langle .,. \rangle_{\mathcal{H}}$.*

The computation of $\mathcal{K}^{perm}$ is expensive because enumerating all permutations requires $|S_d| = d!$. Instead, we propose the following alternative.

The sorting trick. Given the positive semidefinite kernel $\mathcal{K}$, we propose computing a new kernel $\mathcal{K}^{\text{sort}} := \mathcal{K}(\text{sort}\,.,\text{sort}\,.)$, which is also positive semidefinite. We then consider the *sorted* RKHS, $\mathcal{H}^{\text{sort}}$, generated by $\mathcal{K}^{\text{sort}}$. The spaces $\mathcal{H}^{perm}$ and $\mathcal{H}^{\text{sort}}$ are related but not always the same except for special cases where $(\mathcal{H}^{perm})^{\text{sort}} = \mathcal{H}^{perm}$. Note that both $\mathcal{H}^{\text{sort}}$ and $\mathcal{H}^{perm}$ are subspaces of $L^2([0,1]^d, \mathbb{P})$. Let us denote the closure of $\mathcal{H}^{\text{sort}}$ by $\overline{\mathcal{H}}^{\text{sort}}$. We define the $L^2([0,1]^d, \mathbb{P})-$norm and the Euclidean norm in the standard way: $\|f-g\|^2_{L^2([0,1]^d,\mathbb{P})} := \int_{[0,1]^d} |f(t)-g(t)|^2 d\mathbb{P}(t)$ and $\|t_1 - t_2\|_2^2 := \sum_{a=1}^d (t_1^a - t_2^a)^2$. Suppose that we observe n observations of an arbitrary unknown function $f \in \mathcal{H}^{perm}$, in the form $y_i = f(t_i)$ for $i = 1, ..., n$. Let us consider

$$\hat{f} \in \arg \min_{\check{f} \in \mathcal{H}^{\text{sort}}} \|\check{f}\|_{\mathcal{H}^{\text{sort}}}, \quad \text{subject to} \quad \check{f}(t_i) = y_i,\ i = 1, \cdots, n, \tag{5}$$

where $\|.\|_{\mathcal{H}^{\text{sort}}}$ is simply the RKHS norm of $\mathcal{H}^{\text{sort}}$. Because f is permutation invariant, $y_i = f(t_i) = f(\text{sort}\, t_i)$. By the orthogonal projection in $\mathcal{H}$, there exists $\hat{\pi} \in \mathbb{R}^n$ such that $\hat{f} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\pi}_i \mathcal{K}(., \text{sort}\, t_i)$. So naturally we can estimate $f = f \circ \text{sort}$ by $\hat{f} \circ \text{sort} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\pi}_i \mathcal{K}(\text{sort}\,., \text{sort}\, t_i) \in \mathcal{H}^{\text{sort}}$ to ensure that $\hat{f}$ is permutation invariant. We bound the error from approximating f with $\hat{f} \circ \text{sort}$ in the following. This result provides an embedding of $\mathcal{H}^{perm}$ in $\overline{\mathcal{H}}^{\text{sort}}$, inside $L^2([0,1]^d, \mathbb{P})$.

Theorem 5.2 (Approximation Error) *Given a positive semidefinite kernel $\mathcal{K} \in C^{2\nu}([0,1]^d \times [0,1]^d)$ satisfying the invariant property, $\mathcal{K}(.,.) = \mathcal{K}(\sigma.,\sigma.)$ for all $\sigma \in S_d$, then there exists a constant B depending only on the dimension d, such that given any $f \in \mathcal{H}^{perm}$, we have*

$$\left\| f - \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\pi}_i \mathcal{K}(\text{sort}\,., \text{sort}\, t_i) \right\|^2_{L^2([0,1]^d,\mathbb{P})} \leq B \|f\|_{\mathcal{H}} \cdot \sup_{t \in [0,1]^d} \min_{i=1,\cdots,n} \|\text{sort}\, t - \text{sort}\, t_i\|_2^{\nu}.$$

In particular, the right hand side tends to zero as $n \to \infty$ for an appropriately chosen sequence of sets of sample points, hence we have an embedding $\mathcal{H}^{perm} \subset \overline{\mathcal{H}}^{\text{sort}}$ as subspaces of $L^2([0,1]^d, \mathbb{P})$.

The proof of Theorem 5.2 can be found in Section A.8 of the appendix.

This result shows that $\mathcal{K}^{\text{sort}}$ is capable of reproducing anything in $\mathcal{H}^{perm}$. Therefore, when performing a kernel ridge regression (KRR) with the ground truth function known to be in $\mathcal{H}^{perm}$, we may replace $\mathcal{K}^{perm}$ with $\mathcal{K}^{\text{sort}}$. With sorting, one simply takes the kernel function associated with the original RKHS and sorts the inputs.

Extension: kernel ridge regressions with sorting

We now turn to the noisy observation model

$$y_i = f(t_i) + \varepsilon_i,\ i = 1, ..., n, \tag{6}$$

where the noise ε_is are independently drawn from a distribution. Recalling $t_i = (t_i^1, t_i^2, ..., t_i^d)$ for $i = 1, ..., n$, we first sort t_i in the ascending fashion, and in the case of a tie, simply keep the original ordering. Let the sorted vector be denoted by $\text{sort}(t_i)$. We then consider the following estimator:

$$\hat{f} \in \arg \min_{\check{f} \in \mathcal{H}^{\text{sort}}} \frac{1}{2n} \sum_{i=1}^n \left(y_i - \check{f}\left(\text{sort}(t_i)\right) \right)^2 + \lambda \left\| \check{f} \right\|^2_{\mathcal{H}^{\text{sort}}}. \tag{7}$$

Then given t_0, we estimate $f(t_0)$ by $\hat{f}\left(\operatorname{sort} t_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\pi}_{\lambda,i} \mathcal{K}\left(\operatorname{sort} t_0, \operatorname{sort} t_i\right)$, where

$$\hat{\pi}_\lambda = \arg\min_{\pi \in \mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \frac{1}{\sqrt{n}} \sum_{i'=1}^{n} \pi_{i'} \mathcal{K}\left(\operatorname{sort}(t_i), \operatorname{sort}(t_{i'})\right) \right)^2 + \lambda \pi^T \mathbb{K}_{\text{sort}} \pi. \tag{8}$$

In (8), $\mathbb{K}_{\text{sort}} \in \mathbb{R}^{n\times n}$ consist of entries $\frac{1}{n}\mathcal{K}\left(\operatorname{sort} t_i, \operatorname{sort} t_{i'}\right)$.

$\mathcal{K}^{perm}$ vs $\mathcal{K}^{sort}$ and the trade off. In Section 6, we demonstrate numerically that both $\mathcal{K}^{\text{sort}}$ and $\mathcal{K}^{perm}$ generally perform better than $\mathcal{K}$ in KRR for estimating a permutation invariant function, especially for higher d. Compared with $\mathcal{K}^{perm}$ (whose computing time is $d!$), $\mathcal{K}^{sort}$ can be computed with fast sorting of the features (which takes $O(d \log d)$). At the expense of the computational efficiency, a loss of accuracy in using $\mathcal{K}^{\text{sort}}$ over $\mathcal{K}^{perm}$ may arise in some situations but is insignificant in others, as illustrated by our numerical experiments.

6 Simulation studies

All figures and tables that are referred to in this section can be found in Appendix A.9.

6.1 Testing permutation invariance

Throughout we use $N_W = 1000$ to numerically estimate $c_W(\alpha)$, and $N = 1000$ Monte–Carlo replications. To estimate the supremum for T and W, we use the COBYLA maximization algorithm on the smoothened empirical CDF described in Section 2. For key performance indicators, we denote by “Pow" the power of the test (the probability of rejecting H_0 when H_0 is false) and “Cov" the coverage of the test (the probability of not rejecting H_0 when H_0 is true). For the dimension $d = 2$, we demonstrate the performance of our test given a various number of n sample points over $[0, 1]^d$ from the normal distribution

$$\mathcal{N}\left((\mu_1, 0.5), \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & 0.01 \end{pmatrix}\right)$$

modulo $\mathbb{Z}^d$.[4] We adjust (μ_1, σ_1^2) to create various setups for the experiment, with the $(\mu_1, \sigma_1^2) = (0.5, 0.01)$ being the control case where the distribution is perfectly permutation invariant. The results are presented in Table 1.

Next, we study the performance of our test given a fixed $n = 100$ sample points over $[0, 1]^d$ and various higher dimensions $d = 3, 4, 5$, using sample points from the normal distribution:

$$\mathcal{N}\left(\underbrace{(\mu_1, 0.5, \cdots, 0.5)}_{d}, 0.01 \cdot I_d\right)$$

modulo $\mathbb{Z}^d$. The results are presented in Table 2.

Generally, for a fixed d, we can see the improvement in performance with higher n. However, with a fixed n, it becomes increasingly challenging with higher d for the optimization algorithm to estimate the supremum T and W, where a higher n would also be needed.

6.2 Estimating permutation invariant densities

We draw n sample points from $N(0, I_d)$ and compare $\hat{f}(t)$ with $\tilde{f}(t)$ at different values of t when dimensions range from 3 to 5. We use the product of univariate triangular kernels for illustration, although other choices

[4] For example, if a vector $(1.4, 0.7)$ is drawn from $\mathcal{N}\left((\mu_1, 0.5), \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & 0.01 \end{pmatrix}\right)$, then after $\mathbb{Z}^d$ modulo, we produce the sample point $(0.4, 0.7)$ on $[0, 1]^d$.

that satisfy the assumptions in Lemma 3.3(ii) would also work. The simulations in Table 3 have $n = 10000$, $h = 3\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$, and are repeated 1000 times. Figure 2 focuses on $t = (0, 0.25, 0.5, 0.75)$ and exhibits the biases and variances of the standard kernel density estimator $\hat{f}(t)$ and the averaged kernel density estimator $\tilde{f}(t)$ with n increasing from 1000 to 10000. Compared to $\hat{f}(t)$, $\tilde{f}(t)$ exhibits much lower variances and similar biases.

6.3 Kernel ridge regressions of permutation invariant functions

First, we compare the average sample mean squared errors (SMSE) of the sorted KRR (based on the sorted kernels), the permutation invariant KRR (based on the permutation invariant kernels), and the standard KRR (based on the standard kernels) on a $W^{s,2}_{per,perm}([0,1]^d)$ function, $s := \lfloor d/2 + 1 \rfloor$. For any $t \in [0,1]^d$ and $\sigma \in S_d$, $[\sigma(t)]^a$ means the a–th coordinate of the vector $\sigma(t) \in [0,1]^d$. We choose the truth to be a permutation invariant periodic Sobolev function:

$$f(t) := \frac{1}{d!} \sum_{\sigma \in S_d} \left(\sin 2\pi[\sigma(t)]^1 + \cos 6\pi[\sigma(t)]^2 \right).$$

To compute the kernel $\mathcal{K}_{d,s}(t,t')$ according to Lemma A.2 in Appendix A.6, we truncate the infinite series to $|k|_1 = k_1 + \cdots + k_d \leq k_{\max} := 10$. Here, the choice $k_{\max} = 10$ is chosen as it is sufficient to address the highest frequency term in our example function f. We generate $\overline{n} = 500$ observations $\{(t_i, y_i)\}_{i=1}^{\overline{n}}$ with $t_i \sim_{i.i.d.} Unif[0,1]^d$ and $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, 0.1)$ in (6). From the set of observations, we draw a random subsample of size $n = 250$, then perform a 5–fold cross–validation on such subsample to find the average-SMSE-minimizing regularization parameter λ^* for each version of the KRR. We compute the SMSE of each KRR with the corresponding λ^* over the entire subsample, and repeat this process for 100 random subsamples. The average SMSE for various dimensions $d = 2, 3, 4, 5$ are given in Table 4(a).

It is clear that the permutation invariant KRR performs better than the rest because f is a permutation invariant Sobolev function. Nevertheless, the sorted KRR offers an improvement over the standard KRR at higher d, and is much faster to compute than the permutation invariant KRR (in fact, a few minutes vs a few hours on our machine for $d = 5$).

For a (second) fairer comparison, we repeat the simulation above using a different class of kernel, a 1-layer neural network Williams & Barber (1998) for its universal approximation property,

$$\mathcal{K}_{NN}(t,t') := \frac{2}{\pi} \sin^{-1} \frac{2 + 2t^\intercal t'}{\sqrt{(3 + 2t^\intercal t)(3 + 2t'^\intercal t')}}$$

where its permutation invariant and sorted counterparts are also computed. In this case, there is no clear advantage to using the permutation invariant KRR. The results are given in Table 4(b). The sorted KRR performs better than the standard KRR and comparably to the permutation invariant KRR while being much faster to compute.

References

Tarik Bahraoui and Jean-François Quessy. Tests of multivariate copula exchangeability based on lévy measures. *Scandinavian Journal of Statistics*, 49(3):1215–1243, 2022.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. *Advances in neural information processing systems*, 34:18673–18684, 2021.

Benjamin Bloem-Reddy, Yee Whye, et al. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020.

Ziyu Chen, Markos Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Sample complexity of probability divergences under group symmetry. In *International Conference on Machine Learning*, pp. 4713–4734. PMLR, 2023.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786 – 2819, 2013.

Edo Cohen-Karlik, Avichai Ben David, and Amir Globerson. Regularizing towards permutation invariance in recurrent models. *Advances in Neural Information Processing Systems*, 33:18364–18374, 2020.

Michael Harder and Ulrich Stadtmüller. Testing exchangeability of copulas in arbitrary dimension. *Journal of Nonparametric Statistics*, 29(1):40–60, 2017.

Myles Hollander. A nonparametric test for bivariate symmetry. *Biometrika*, 58(1):203–212, 1971.

Jan Kalina and Patrik Janáček. Testing exchangeability of multivariate distributions. *Journal of Applied Statistics*, 50(15):3142–3156, 2023.

Stefan Klus, Patrick Gelß, Feliks Nüske, and Frank Noé. Symmetric and antisymmetric kernels for machine learning problems in quantum physics and chemistry. *Machine Learning: Science and Technology*, 2(4): 045016, 2021.

Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Arun Kumar Kuchibhotla. Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*, 2020.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Guanjie Lyu and Mohamed Belalia. Testing symmetry for bivariate copulas using bernstein polynomials. *Statistics and Computing*, 33(6):128, 2023.

Erich Novak, Mario Ullrich, Henryk Woźniakowski, and Shun Zhang. Reproducing kernels of sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability. *Analysis and Applications*, 16(05):693–715, 2018.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pp. 1094–1103. PMLR, 2017.

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. *Advances in Neural Information Processing Systems*, 36, 2023.

Yujin Tang and David Ha. The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22574–22587, 2021.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 1 edition, 2009.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1342–1351, 1998.

Takemi Yanagimoto and Masaaki Sibuya. Test of symmetry of a bivariate distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 105–115, 1976.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

A Appendix

A.1 Proof of Theorem 2.2

Proof: Consider a sample $\{t_i\}_{i=1,\cdots,n}$ drawn from a distribution with a continuous CDF F. Let us choose $\{z_k \in [0,1]\}_{k=1\cdots,n^m}$ for some $m \geq 4$ such that each slab

$$B_k^a := \{t \in [0,1]^d \mid z_{k-1} \leq t^a \leq z_k\}$$

contains equal probability for any coordinate $a = 1,\cdots,d$ and $k = 1,\cdots,n^m$: i.e.

$$\int_{B_k^a} 1 \cdot dF(t) = \frac{1}{n^m},$$

where we take $z_0 = 0$ by convention, and let us define: $\{v_{k_1,\cdots,k_p} := (z_{k_1},\cdots,z_{k_p}) \in [0,1]^p\}_{k_1,\cdots,k_d=1,\cdots,n^m}$. Fix a set $A \subset [0,1]^d$ bounded away from the boundaries and the diagonal of $[0,1]^d$ such that

$$\inf_{v\in A} \int_{[0,1]^d} \left([t \leq \operatorname{sort} v] - [t \leq v]\right)^2 dF(t) > 0$$

and let $\{\widetilde{v}_j \subset A\}_{j=1,\cdots,n^{md}}$ be any subset of n^{md} points in A. For convenience, let us relabel the elements of $\{v_{k_1,\cdots,k_d}\} \cup \{\widetilde{v}_j\}$ in some ways as $\{v_j\}_{j=1,\cdots,2n^{md}}$ where $v_j \in \{v_{i_1,\cdots,i_d}\}$ for $j = 1,\cdots,n^{md}$, and $v_j \in \{\widetilde{v}_j\}_{j=1,\cdots,n^{md}}$ for $j = n^{md}+1,\cdots,2n^{md}$. For $j = 1,\cdots,n^{md}$, define the box

$$B_j := \{t \in [0,1]^d \mid t \leq v_j \text{ and } t \not\leq v_{j'} \text{ for any } v_{j'} < v_j\},$$

and let us assume that $\int_{B_j} 1 \cdot dF(t) \leq \frac{1}{n^{(m-1)d}}$ [5] Alternatively, we can write $B_j = B_{k_1,\cdots,k_d} := \bigcap_{a=1}^d B_{k_a}^a$ if $v_j = v_{k_1,\cdots,k_d}$. We apply the results of Chernozhukov et al. (2013) with

$$x_{ij} := \begin{cases} [t_i \leq \operatorname{sort} v_j] - [t_i \leq v_j], & i = 1,\cdots,\lceil n/2 \rceil - 1; j = 1,\cdots,2n^{md} \\ [t_i \leq \operatorname{sort} v_{j+n^{md}}] - [t_i \leq v_{j+n^{md}}], & i = \lceil n/2 \rceil,\cdots,n; j = 1,\cdots,n^{md} \\ [t_i \leq \operatorname{sort} v_{j-n^{md}}] - [t_i \leq v_{j-n^{md}}], & i = \lceil n/2 \rceil,\cdots,n; j = n^{md}+1,\cdots,2n^{md} \\ -x_{i,j-2n^{md}}, & \forall i; j = 2n^{md}+1,\cdots,4n^{md} \end{cases}.$$

Then $x_i = (x_{i1},\cdots,x_{i,2n^{md}}) \in [0,1]^{2n^{md}}$ for $i = 1,\cdots,n$ are independent random vectors. Moreover, it follows from the permutation invariance of the CDF F that each x_i is centered, i.e. $\mathbb{E}[x_{ij}] = 0$. For convenience, we introduce the following notation

$$Z_n(t;\{c_i\}) := \frac{1}{\sqrt{n}} \left| \sum_{i=1}^n \left([t_i \leq \operatorname{sort} v_j] - [t_i \leq v_j]\right) c_i \right|$$

and

$$T_0 := \max_{1\leq j\leq 4n^{md}} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} = \max_{1\leq j\leq 2n^{md}} Z_n(v_j;\{c_i = 1\})$$

$$W_0 := \max_{1\leq j\leq 4n^{md}} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_j = \max_{1\leq j\leq 2n^{md}} \frac{1}{\sqrt{n}} Z_n(v_j;\{e_i\}), \quad e_i \sim \mathcal{N}(0,1).$$

[5] If necessary, we may start the construction from finding $\{\widetilde{z}_k^a\}_{k=1,\cdots,2n^{m-1};a=1,\cdots,d}$, $\widetilde{z}_{i_1,\cdots,i_d} := (\widetilde{z}_{i_1}^1,\cdots,\widetilde{z}_{i_d}^d) \in [0,1]^d$, such that each $\widetilde{B}_{i_1,\cdots,i_d} := \{t \in [0,1]^d \mid \widetilde{z}_{i_1-1,\cdots,i_d-1} \leq t \leq \widetilde{z}_{i_1,\cdots,i_d}\}$ contains equal probability of $\frac{1}{2^d n^{(m-1)d}}$. Hence each slab $\widetilde{B}_k^a := \{t \in [0,1]^d \mid \widetilde{z}_{k-1}^a \leq t^a \leq \widetilde{z}_k^a\}$ contains equal probability of $\frac{1}{2n^{m-1}}$. The needed symmetric grid $\{z_k\}_{k=1,\cdots,n^m}$ must be 'smaller' than each $\{\widetilde{z}_k^a\}$ in a sense that for any $a = 1,\cdots,p$ we have $[z_{k-1}, z_k] \subset [\widetilde{z}_{k'-1}^a, \widetilde{z}_{k'+1}^a]$ for some $k' \in \{1,\cdots,2n^{m-1}-1\}$, which implies $\mathbb{P}[B_j] \leq 2^d \cdot \frac{1}{2^d n^{(m-1)d}} = \frac{1}{n^{(m-1)d}}$.

When there is no ambiguity, we will adopt the notation $Z_n(t) := Z_n(t; \{c_i = 1\})$. We would like to compare T_0 and W_0 to

$$T := \sup_{t\in[0,1]^d} Z_n(t), \quad W := \sup_{t\in[0,1]^d} Z_n(t; \{e_i\}).$$

More precisely, as required in Chernozhukov et al. (2013), we will show that there exists $\zeta_1, \zeta_2 \geq 0$ both depending on n such that $\zeta_1\sqrt{\log 4n^{md}} + \zeta_2 \leq C_2 n^{-c_2} \to 0$ as $n \to \infty$ for some constants $C_2, c_2 > 0$, and that

$$\mathbb{P}\left[|T - T_0| > \zeta_1\right] < \zeta_2, \quad \mathbb{P}\left[\mathbb{P}_e\left[|W - W_0| > \zeta_1\right] > \zeta_2\right] < \zeta_2. \tag{9}$$

For any fixed drawn $\{e_i\}_{i=1,\cdots,n}$, the supremum of $Z_n(t; \{e_i\})$ can be reached at some unsorted point t^* of the form

$$t^* = \left(t_{i_1}^{a_1}, \cdots, t_{i_d}^{a_d}\right)$$

for some $a_1, \cdots, a_d \in \{1, \cdots, d\}$ and $i_1, \cdots, i_d \in \{1, \cdots, n\}$. Suppose that $t^* \in B_{j^*} = \bigcap_{a=1}^p B^a_{i^*_a}$ for some $j^* \in \{1, \cdots, n^{md}\}$. If $\text{sort}\, v_{j^*} =: \pi^*(v_{j^*}) \neq v_{j^*}$ i.e. t^* is not too close to the diagonal, and $\bigcup_{a=1}^d B^a_{i^*_a} \cup \bigcup_{a=1}^d B^{\pi^*(a)}_{i^*_a}$ contains no other sampled points apart from $t_{i_1}, \cdots, t_{i_d}$, then

$$Z_n(x_{j^*}; \{e_i\}) = Z_n(t^*; \{e_i\}) = \sup_{t\in[0,1]^d} Z_n(t; \{e_i\}).$$

A sufficient condition for the above requirements to be satisfied is that each interval $[z_{k-1}, z_k]$, $k = 1, \cdots, n^m$ contains at most $p-1$ coordinates of at most one sampled vectors $\{t_{i'}\}_{i'=1,\cdots,n}$, i.e. $\forall i = 1, \cdots, n^m$:

$$\begin{aligned} p - 1 \geq \left|[z_{i-1}, z_i] \cap \{t^1_{i'}, \cdots, t^d_{i'}\}\right| &\geq 0, \quad \forall i' = 1, \cdots, n \\ \left|[z_{i-1}, z_i] \cap \{t^1_{i'}, \cdots, t^d_{i'}\}\right| &> 0, \quad \text{for at most one } i' = 1, \cdots, n. \end{aligned}$$

If this is satisfied, we say that there is no *coordinate collision*. We compute the upper bound for the collision probability as follows. Let $I \subset \{1, \cdots, n^m\}$ be any fixed subset of size $(n-1)d$. Then,

$$\begin{aligned} \mathbb{P}\left[\text{Coordinate collision}\right] &\leq \mathbb{P}\left[\bigcup_{i=1}^n \left\{ t_i \in \bigcup_{i\in\{1,\cdots,n^m\}} B_{i,\cdots,i} \cup \bigcup_{i\in I, a\in\{1,\cdots,d\}} B^a_i \right\}\right] \\ &\leq n \cdot n^m \cdot \frac{1}{n^{(m-1)d}} + n \cdot d \cdot n(d-1) \cdot \frac{1}{n^m} \sim \frac{1}{n^{md-d-m-1}} + \frac{d^2}{n^{m-2}} \to 0, \end{aligned}$$

as $n \to \infty$, since $md - d - m - 1 > 0$ for all $m \geq 4, d \geq 2$ and $d^2/n^{m-2} \lesssim 1/n^{m-2-2/7} \to 0$ given that $d = o(n^{1/7})$. In the second inequality, the first term came from counting the number of diagonal boxes, each of them has probability bounded above by $\frac{1}{n^{(m-1)d}}$ by our construction. The second term counts the upper bound for the probability that at least one of the coordinates of any t_i shares the interval with one of the coordinates of one of the other $n-1$ drawn vectors. Since each of the $n(d-1)$ slabs contains an equal probability of $\frac{1}{n^m}$, the exact choice of I does not matter, as long as $|I| = (n-1)d$. It follows that

$$\mathbb{P}\left[|T - T_0| > 0\right] < \frac{1}{n^{md-d-m-1}} + \frac{d^2}{n^{m-2}}$$

and

$$\mathbb{P}\left[\mathbb{P}_e\left[|W - W_0| > 0\right] > \frac{1}{n^{mp-p-m-1}} + \frac{d^2}{n^{m-2}}\right] \leq \mathbb{P}\left[\mathbb{P}_e\left[|W - W_0| > 0\right] > 0\right] < \frac{1}{n^{md-d-m-1}} + \frac{d^2}{n^{m-2}}.$$

Therefore, (9) holds with $\zeta_1 := 0, \zeta_2 := \frac{1}{n^{md-d-m-1}} + \frac{d^2}{n^{m-2}}$

In the language of Chernozhukov et al. (2013), we also have for $j = 1, \cdots, 4n^{mp}$:

$$\bar{E}[x^2_{ij}] := \frac{1}{n}\sum_{i=1}^n E[x^2_{ij}] \geq \frac{1}{n}\left\lfloor\frac{n}{2}\right\rfloor \int_{[0,1]^d} \left([t \leq \text{sort}\, \widetilde{v}_{j'}] - [t \leq \widetilde{v}_{j'}]\right)^2 dF(t)$$

$$+\frac{1}{n}\left\lfloor\frac{n}{2}\right\rfloor\int_{[0,1]^p}\left([t\leq \mathrm{sort}\, v_{i_1,\cdots,i_p}]-[t\leq v_{i_1,\cdots,i_p}]\right)^2 dF(t)$$
$$\geq \frac{1}{3}\inf_{v\in A}\int_{[0,1]^p}\left([t\leq \mathrm{sort}\, v]-[t\leq v]\right)^2 dF(t) =: c_1 > 0$$

where the first inequality holds for some $j'\in\{1,\cdots,n^{mp}\}$ and $(i_1,\cdots,i_p)\in\{1,\cdots,n^m\}^p$ depending on j. On the other hand, it is clear that $\bar{E}[x_{ij}^2]\leq C_1 := 1$. We can also choose $B_n = 1$ to satisfy the condition:

$$\max_{k=1,2}\bar{E}\left[|x_{ij}|^{2+k}/B_n^k\right]+E[\exp\left(|x_{ij}|/B_n\right)]\leq 1+e\leq 4.$$

It follows from Corollary 3.1 of Chernozhukov et al. (2013) and condition E.1. with

$$\frac{B_n^2(\log(2n^{mp}\cdot n))^7}{n}\leq\frac{\left[(mp+1)\log n+\log 2\right]^7}{n}\leq C_2 n^{-c_2} \tag{10}$$

for some $C_2, c_2 > 0$, since we have assumed $p = o(n^{1/7})$, that there exists $c > 0, C > 0$ depending only on C_1, c_1, C_2, c_2 such that

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left[T > c_W(\alpha)\right]-\alpha\right|\leq Cn^{-c}.$$

Unpacking the definition of T, we find that this is the statement of the theorem. □

A.2 Proof of Lemma 3.2

Proof: The expected value of a standard kernel estimator is computed as

$$\mathbb{E}[\hat{f}(t)] = f(t)+\frac{h^2}{2}tr\left(\frac{\partial^2 f(t)}{\partial t\partial t^T}\int vv^T K(v)dv\right)+\text{higher order terms.}$$

For a permutation σ, we have that

$$\mathbb{E}[\hat{f}(\sigma(t))] = f(\sigma(t))+\frac{h^2}{2}tr\left(\frac{\partial^2 f(\sigma(t)}{\partial t\partial t^T}\int vv^T K(v)dv\right)\text{higher order terms.}$$

Because f is permutation invariant, $f(\sigma(t)) = f(t)$ and $\frac{\partial^2 f(\sigma(t)}{\partial t\partial t^T}=\frac{\partial^2 f(t)}{\partial t\partial t^T}$. □

A.3 Proof of Lemma 3.3

Proof: The first variance is a known result for multivariate kernel density estimators. To obtain the second variance, we have

$$\begin{aligned}
\mathbb{E}(\tilde{f}(t)^2) &= \frac{1}{(\bar{d})^2}\mathbb{E}\Big(\sum_{\sigma\in S_d^*}\hat{f}(\sigma(t))\Big)^2\\
&= \frac{1}{(\bar{d})^2}\mathbb{E}\left(\sum_{\sigma\in S_d^*}\hat{f}(\sigma(t))^2+\sum_{\sigma\neq\pi}\hat{f}(\sigma(t))\hat{f}(\pi(t))\right)\\
&= \frac{1}{(\bar{d})^2}\sum_{\sigma\in S_d^*}\mathbb{E}\Big(\hat{f}(\sigma(t))^2\Big)+\frac{1}{(\bar{d})^2}\sum_{\sigma\neq\pi}\mathbb{E}\Big(\hat{f}(\sigma(t))\hat{f}(\pi(t))\Big).
\end{aligned}$$

In terms of the first term, we have

$$\mathbb{E}\left(\hat{f}(\sigma(t))^2\right) = \frac{1}{nh^d}f(\sigma(t))\int K(v)^2 dv+\text{higher order terms.}$$

Since there are $\bar{d}$ permutations in S_d^*, we have

$$\mathbb{E}[\tilde{f}(t)^2] = \frac{1}{\bar{d}}\frac{1}{nh^d}f(t)\int K(v)^2 dv + \frac{1}{(\bar{d})^2}\sum_{\sigma\neq\pi}\mathbb{E}\Big(\hat{f}(\sigma(t))\hat{f}(\pi(t))\Big) + \text{higher order terms.}$$

We now turn to the cross-product terms.

$$\begin{aligned}
\mathbb{E}\Big(\hat{f}(\sigma(t))\hat{f}(\pi(t))\Big) &= \mathbb{E}\Big(\frac{1}{nh^d}\Big(\sum_{i=1}^n K\Big(\frac{\sigma(t)-t^i}{h}\Big)\Big)\frac{1}{nh^d}\Big(\sum_{i=1}^n K\Big(\frac{\pi(t)-t^i}{h}\Big)\Big)\Big) \\
&= \frac{1}{n^2h^{2d}}\mathbb{E}\left(\sum_{i=1}^n K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big) + \sum_{i\neq j}K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) \\
&= \frac{1}{n^2h^{2d}}\sum_{i=1}^n \mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right) + \frac{1}{n^2h^{2d}}\sum_{i\neq j}\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) \\
&= \frac{1}{nh^{2d}}\underbrace{\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right)}_{A} + \frac{n-1}{nh^{2d}}\underbrace{\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right)}_{B}.
\end{aligned}$$

We start with the second term, B:

$$\begin{aligned}
\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) &= \int K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)f(t^i)f(t^j)dt^i dt^j \\
&= \left(\int K\Big(\frac{\sigma(t)-t^i}{h}\Big)f(t^i)dt^i\right)\left(\int K\Big(\frac{\pi(t)-t^i}{h}\Big)f(t^i)dt^i\right).
\end{aligned}$$

These are the same integrals except for the difference of σ and π. Without loss of generality, we work with the first one:

$$\begin{aligned}
\int K\Big(\frac{\sigma(t)-t^i}{h}\Big)f(t^i)dt^i &= h^d\int K(u)f(hu+\sigma(t))du \\
&\approx h^d\int\Big[K(u)f(\sigma(t)) + h\underbrace{K(u)u^T}_{\text{Integrates to }0}\nabla f(\sigma(t))\Big]du \\
&= h^d f(\sigma(t)).
\end{aligned}$$

There are n^2-n terms where $i\neq j$ which means that

$$\frac{1}{n^2h^{2d}}\mathbb{E}\left(\sum_{i\neq j}K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) \approx \frac{n-1}{n}f(\sigma(t))f(\pi(t)) = f(t)^2 - \frac{1}{n}f(t)^2.$$

We now turn to the first term, A:

$$\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right) = \int K\Big(\frac{\sigma(t)-\pi(t)}{h} + \frac{\pi(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)f(t^i)dt^i.$$

Letting $u = \frac{t^i-\pi(t)}{h}$, we have

$$\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right) = h^d\int K\Big(\frac{\sigma(t)-\pi(t)}{h} - u\Big)K\Big(u\Big)f(hu+\pi(t))du$$

$$\approx h^d \int K\Big(\frac{\sigma(t)-\pi(t)}{h}-u\Big)K\Big(u\Big)f(\pi(t))du$$
$$= h^d f(\pi(t))\int K\Big(\frac{\sigma(t)-\pi(t)}{h}-u\Big)K\Big(u\Big)du$$
$$= h^d f(t)(K*K)\Big(\frac{\sigma(t)-\pi(t)}{h}\Big).$$

Putting everything together yields the second variance.

For the results in part (ii), note that we have

$$K*K\Big(\frac{\sigma(t)-\pi(t)}{h}\Big)=\prod_{j=1}^{d}(k*k)\Big(\frac{\sigma_j(t)-\pi_j(t)}{h}\Big).$$

The following observations can be made about the term above:

1. given the conditions on $k(v)$ in part (ii) of Lemma 3.3, then $(k*k)\Big(\frac{\sigma_j(t)-\pi_j(t)}{h}\Big)$ is *strictly* smaller than $\int k(v)^2dv$ when $\sigma_j(t)\neq\pi_j(t)$;
2. given the observation above and that $\sigma,\pi\in S_d^*$, by the construction of S_d^*, we have

$$\prod_{j=1}^{d}k*k\Big(\frac{\sigma_j(t)-\pi_j(t)}{h}\Big)\leq b^{\bar{d}}\int K(v)^2dv$$

   where $0<b<1$ depends on $\frac{\sigma(t)-\pi(t)}{h}$.

□

A.4 Proof of Theorem 4.2

Proof: We use the argument in Kolmogorov & Tikhomirov (1959) and Lemma A.1. When permutation invariance is absent, to derive an upper bound on $\log N_\infty(\delta,\mathcal{U})$, we consider a $b_{d,\gamma}^{-1}\delta^{\frac{1}{\gamma+1}}$ −grid of points (where $b_{d,\gamma}$ is a function of (d,γ) only and independent of δ) in each dimension of $[0,1]^d$:

$$x_{0,j}=0<x_{1,j}<\cdots<x_{s-1,j}<x_{s,j},\quad j\in\{1,...,d\}$$

with $s\lesssim b_{d,\gamma}\delta^{\frac{-1}{\gamma+1}}$, and show that bounding $N_\infty(\delta,\mathcal{U})$ can be reduced to bounding the cardinality of

$$\Lambda=\left\{\left(\left\lfloor\frac{D^pf(x_{i_1,1},...,x_{i_d,d})}{\delta_k}\right\rfloor,0\leq i_1,...,i_d\leq s,\,0\leq k\leq\gamma\right):f\in\mathcal{U}\right\}$$

with $\lfloor x\rfloor$ denoting the largest integer smaller than or equal to x. Then, using the fact that $D^pf(0)=0$ for all p with $P=k$ $(k=0,...,\gamma)$, the argument in Kolmogorov & Tikhomirov (1959) implies that $|\Lambda|\leq c^{s^d}$, where $c\in(0,\infty)$ is a constant independent of δ and (d,γ). Now with permutation invariance, by Lemma A.1, the number of points we need to consider scales as $\frac{1}{d!}s^d$. This fact is also applied along with the construction of the class of functions in Kolmogorov & Tikhomirov (1959) and the relationship between covering numbers and packing numbers to yield the lower bound. In addition, $\log N_2(\delta,\mathcal{U})\lesssim\log N_\infty(\delta,\mathcal{U})$. Standard argument in the literature using the Vasharmov-Gilbert Lemma and the relationship between covering numbers and packing numbers further give $\log N_2(\delta,\mathcal{U})\gtrsim\log N_\infty(\delta,\mathcal{U})$. In sum, $\log N_2(\delta,\mathcal{U})\asymp\log N_\infty(\delta,\mathcal{U})\asymp b_{d,\gamma}^d\delta^{\frac{-d}{\gamma+1}}$ and $\log N_2(\delta,\mathcal{U}^{perm})\asymp\log N_\infty(\delta,\mathcal{U}^{perm})\asymp\frac{1}{d!}b_{d,\gamma}^d\delta^{\frac{-d}{\gamma+1}}$.

Lemma A.1 *Let* $\mathcal{P}_d^b=\left\{x\in[0,b]^d:0\leq x_1\leq x_2\leq...\leq x_d\leq b\right\}$. *Then the volume of* $\mathcal{P}_d^b$, $Vol(\mathcal{P}_d^b)=\frac{b^d}{d!}$.

Proof: We show Lemma A.1 by induction.

Base case: If $d=1$, $\mathcal{P}_d^b = [0,b]$. Then, $\mathrm{Vol}(\mathcal{P}_1^b) = \frac{b}{1!}$.

Inductive step: Suppose that $\mathrm{Vol}(\mathcal{P}_d^b) = \frac{b^d}{d!}$. Then,

$$\mathrm{Vol}(\mathcal{P}_{d+1}^b) = \int_{x \in \mathcal{P}_{d+1}^b} dx = \int_0^b dx_{d+1} \underbrace{\int_0^{x_{d+1}} dx_d \cdots \int_0^{x_2} dx_1}_{\mathrm{Vol}(\mathcal{P}_d^{x_{d+1}})} = \int_0^b \frac{x_{d+1}^d}{d!} dx_{d+1} = \frac{b^{d+1}}{(d+1)!}.$$

□

A.5 Proof of Theorem 4.4

Proof: From definition, we have $\mu_k \to 0$ as $\max_{a=1,\cdots,d} k^a \to \infty$, therefore we can find $\mathbb{K}_\delta := \left\{k \in \mathbb{Z}_{\geq 0}^d \mid \max_{a=1,\cdots,d} k^a \leq \bar{k}\right\}$ for some $\bar{k} > 0$ such that $\sum_{k \in \mathbb{Z}_{\geq 0}^d \setminus \mathbb{K}_\delta} \beta_k^2 \leq \delta^2$ for all $\beta \in \mathcal{E}$. Then any δ–cover $\{\beta^1, \cdots, \beta^N\}$ of the $D_\delta := |\mathbb{K}_\delta|$–dimensional truncated ellipsoid $\widetilde{\mathcal{E}} := \{\beta \in \mathcal{E} \mid \beta_k = 0, \forall k \notin \mathbb{K}_\delta\}$ is a $\sqrt{2}\delta$–cover of $\mathcal{E}$, since

$$\min_{j=1,\cdots,N} \|\beta - \beta^j\|_{l^2}^2 = \min_{j=1,\cdots,N} \sum_{k \in \mathbb{K}_\delta} (\beta_k - \beta_k^j)^2 + \sum_{k \in \mathbb{Z}_{\geq 0}^d \setminus \mathbb{K}_\delta} (\beta_k)^2 \leq 2\delta^2.$$

for any $\beta \in \mathcal{E}$. It follows from Lemma 5.7 Wainwright (2019) that

$$\left(\frac{\sqrt{2}}{\delta}\right)^{D_\delta} \frac{vol\left(\widetilde{\mathcal{E}}\right)}{vol\left(\mathbb{B}_2^{D_\delta}(1)\right)} \leq N(\delta, \mathcal{E}, \|.\|_{l_2}) \leq \left(\frac{2\sqrt{2}}{\delta}\right)^{D_\delta} \frac{vol\left(\widetilde{\mathcal{E}} + \mathbb{B}_2^{D_\delta}(\delta/2)\right)}{vol\left(\mathbb{B}_2^{D_\delta}(1)\right)}.$$

Let $\underline{\mu}_\delta := \min_{k \in \mathbb{K}_\delta} \mu_k$ and $\bar{\mu}_\delta := \max_{k \in \mathbb{K}_\delta} \mu_k$, then it follows from $\mathbb{B}_2^{D_\delta}(\sqrt{\underline{\mu}_\delta}) \subset \widetilde{\mathcal{E}} \subset \mathbb{B}_2^{D_\delta}(\sqrt{\bar{\mu}_\delta})$ that:

$$\left(\frac{\sqrt{2\underline{\mu}_\delta}}{\delta}\right)^{D_\delta} \leq N(\delta, \mathcal{E}, \|.\|_{l_2}) \leq \left(\frac{2\sqrt{2\bar{\mu}_\delta}}{\delta} + 1\right)^{D_\delta}.$$

On the other hand, let $\mathbb{K}_\delta$ be as chosen previously, and let $\mathbb{K}_\delta^{perm} := \mathrm{sort}\, \mathbb{K}_\delta$ then $D_\delta^{perm} := |\mathbb{K}_\delta^{perm}| \asymp \frac{D_\delta}{d!}$ by the construction that $\mathbb{K}_\delta$ contains the entire S_d–orbit of all its elements. Then $\sum_{k \in \mathrm{sort}\, \mathbb{Z}_{\geq 0}^d \setminus \mathbb{K}_\delta^{perm}} \beta_k^2 \leq \sum_{k \in \mathbb{Z}_{\geq 0}^d \setminus \mathbb{K}_\delta} \beta_k^2 \leq \delta^2$, and therefore any δ–cover $\{\beta'^1, \cdots, \beta'^{N'}\}$ of the D_δ^{perm}–dimensional truncated permutation invariant ellipsoid $\widetilde{\mathcal{E}}^{perm} := \{\beta \in \mathcal{E}^{perm} \mid \beta_k = 0, \forall k \notin \mathbb{K}_\delta^{perm}\}$ is a $\sqrt{2}\delta$–cover of $\mathcal{E}^{perm}$, since

$$\min_{j=1,\cdots,N'} \|\beta - \beta'^j\|_{l^2}^{perm2} = \min_{j=1,\cdots,N'} \sum_{k \in \mathbb{K}_\delta^{perm}} (\beta_k - \beta_k'^j)^2 + \sum_{k \in \mathrm{sort}\, \mathbb{Z}_{\geq 0}^d \setminus \mathbb{K}_\delta^{perm}} (\beta_k)^2 \leq 2\delta^2$$

for any $\beta \in \mathcal{E}^{perm}$. Following the same analysis as before, we obtain

$$\left(\frac{\sqrt{2\underline{\mu}_\delta}}{\delta}\right)^{D_\delta^{perm}} \leq N(\delta, \mathcal{E}^{perm}, \|.\|_{l_2}^{perm}) \leq \left(\frac{2\sqrt{2\bar{\mu}_\delta}}{\delta} + 1\right)^{D_\delta^{perm}}.$$

The result follows by identifying $\underline{g}(\delta) := D_\delta \log\left(\frac{\sqrt{2\underline{\mu}_\delta}}{\delta}\right)$ and $\bar{g}(\delta) := D_\delta \log\left(\frac{2\sqrt{2\bar{\mu}_\delta}}{\delta} + 1\right)$. □

A.6 Lemma A.2 and its proof

Here, we derive the reproducing kernel of the space of periodic Sobolev functions $W^{s,2}_{per}([0,1]^d)$, i.e. the subspace of the Sobolev space $W^{s,2}([0,1]^d)$ for functions f such that $f(t+k) = f(t)$ for any $k \in \mathbb{Z}^d$. We only consider the case when $s > d/2$ so that we have an embedding $W^{s,2}([0,1]^d) \subset C^{0,s-d/2}([0,1]^d)$ to guarantee that $W^{s,2}_{per}([0,1]^d)$ is an RKHS. For reasons that are not obvious to us, results on reproducing kernels for *multivariate* Sobolev space appear sparse in the literature despite its importance. One result we found is for the reproducing kernel of $W^{s,2}(\mathbb{R}^d)$ and derived in Novak et al. (2018). Our derivation supplements their result for the compact domain case. The periodicity is a natural condition that is needed to make sense of the specified smoothness at the boundary of $[0,1]^d$. Equivalently, we may consider Sobolev functions on a d–dimensional torus $\mathbb{T}^d$.

Lemma A.2 *The reproducing kernel $\mathcal{K} : [0,1]^d \times [0,1]^d \to \mathbb{R}$ for the periodic Sobolev space $W^{s,2}_{per}([0,1]^d)$ with $s > d/2$ is*

$$\mathcal{K}_{d,s}(t,t') = \sum_{k \in \mathbb{Z}^d_{\geq 0}} \frac{2\cos 2\pi k \cdot (t-t')}{v_{d,s}[k]^2} - 1,$$

where $v_{d,s}[k] := \left[\sum_{|\alpha|\leq s} \prod_{j=1}^d (2\pi k_j)^{2\alpha_j}\right]^{1/2}$ and the corresponding eigenvalues $\{\mu_k = 1/v_{d,s}[k]^2\}_{k\in\mathbb{Z}^d}$.

By restricting to real–valued functions, the eigenvectors of the corresponding Hilbert-Schmidt operator are $\{e^+_k := t \mapsto \cos 2\pi k \cdot t, e^-_k := t \mapsto \sin 2\pi k \cdot t\}_{k\in\mathbb{Z}^d_{\geq 0}}$ with eigenvalues $\{\mu_k = 1/v_{d,s}[k]^2\}_{k\in\mathbb{Z}^d_{\geq 0}}$, and the RKHS can be written as an ellipsoid taking the same form as in Definition 4.3:

$$W^{s,2}_{per}([0,1]^d) \cong \left\{ (\beta_k)_{k\in\mathbb{Z}^d_{\geq 0}} \mid \sum_{k\in\mathbb{Z}^d_{\geq 0}} v_{d,s}[k]^2 \beta_k^2 < \infty \right\},$$

where $\langle f, g\rangle_{W^{s,2}_{per}} = \sum_{k\in\mathbb{Z}^d_{\geq 0}} v_{d,s}[k]^2 \beta_{f,k}\beta_{g,k}$. For the space of permutation invariant periodic Sobolev functions $W^{s,2}_{per,perm}([0,1]^d) \subset W^{s,2}_{per}([0,1]^d)$, we simply restrict the above ellipsoid to $\beta_{\text{sort}\, k} = \beta_k$.

Proof: For any $f \in W^{s,2}_{per}([0,1]^d) \subset L^2([0,1]^d)$, the corresponding Fourier series is given by

$$f(t) = \sum_{k_1,\cdots,k_d=-\infty}^{+\infty} \mathcal{F}f[k] e^{i2\pi k\cdot t}$$

where

$$\mathcal{F}f[k] := \int_{[0,1]^p} f(t) e^{-i2\pi k\cdot t} dt.$$

The inner products are given by:

$$\langle f, g\rangle_{L^2} := \int_{[0,1]^p} f(t)^* g(t) dt, \quad \langle f, g\rangle_{W^{s,2}} := \sum_{|\alpha|\leq s} \langle D^\alpha f, D^\alpha g\rangle_{L^2}.$$

The periodic properties help us easily compute the Fourier coefficients of any derivatives:

$$\mathcal{F}(D^\alpha f)[k] = \prod_{j=1}^d (i2\pi k_j)^{\alpha_j} \mathcal{F}f[k].$$

For convenience, we define $v_{d,s}[k] := \left[\sum_{|\alpha|\leq s} \prod_{j=1}^d (2\pi k_j)^{2\alpha_j}\right]^{1/2}$. From Parseval's Theorem we also know that

$$\langle f, g\rangle_{L^2} = \sum_{k_1,\cdots,k_d=0}^{\infty} (\mathcal{F}f[k])^* (\mathcal{F}g[k]) =: \langle \mathcal{F}f, \mathcal{F}g\rangle_{l^2}.$$

Therefore,

$$\langle f,g\rangle_{W^{s,2}_{per}} = \sum_{|\alpha|\le s}\langle D^\alpha f, D^\alpha g\rangle_{L^2} = \sum_{|\alpha|\le s}\langle \mathcal{F}D^\alpha f, \mathcal{F}D^\alpha g\rangle_{l^2}$$

$$= \sum_{|\alpha|\le s}\sum_{k_1,\cdots,k_d=-\infty}^{+\infty}\prod_{j=1}^{p}(2\pi k_j)^{2\alpha_j}\mathcal{F}f[k]^*\mathcal{F}g[k] = \langle v_{d,s}\mathcal{F}f, v_{d,s}\mathcal{F}g\rangle_{l^2}\,.$$

Let $K_{p,s}$ be the reproducing kernel. We have $K_{p,s}(.,t)\in W^{s,2}_{per}([0,1]^d)$ and for any $f\in W^{s,2}_{per}([0,1]^d)$:

$$f(t) = \langle f, K_{d,s}(.,t)\rangle_{W^{s,2}_{per}} = \langle v_{d,s}\mathcal{F}f, v_{d,s}\mathcal{F}K_{d,s}(.,t)\rangle_{l^2}\,.$$

On the other hand,

$$f(t) = \mathcal{F}^{-1}\mathcal{F}f(t) = \sum_{k_1,\cdots,k_p=-\infty}^{+\infty}\mathcal{F}f[k]e^{-i2\pi k\cdot t}$$

$$= \sum_{k_1,\cdots,k_d=-\infty}^{+\infty} v_{d,s}[k]\mathcal{F}f[k]^*\cdot v_{d,s}[k]\frac{e^{-i2\pi k\cdot t}}{v_{d,s}[k]^2} = \left\langle v_{d,s}\mathcal{F}f, v_{d,s}\frac{e^{-i2\pi k\cdot t}}{v^2_{d,s}}\right\rangle_{l^2}\,.$$

Comparing the two above, since f is arbitrary, we must have $\mathcal{F}K_{d,s}(.,t)[k] = \frac{e^{-i2\pi k\cdot t}}{v_{d,s}[k]^2}$, which means:

$$K_{d,s}(t,t') = \sum_{k_1,\cdots,k_d=-\infty}^{+\infty}\frac{e^{i2\pi k\cdot(t-t')}}{v_{d,s}[k]^2} = \sum_{k\in\mathbb{Z}^d_{\ge 0}}\frac{2\cos 2\pi k\cdot(t-t')}{v_{d,s}[k]^2} - 1.$$

The corresponding Hilbert-Schmidt operator is $T_K : L^2([0,1]^d)\to L^2([0,1]^d)$:

$$T_K[f](t) := \int_{[0,1]^d} K(t,t')f(t')dt'$$

with eigenvectors $\{e_k := t\mapsto e^{i2\pi k\cdot t}\}_{k\in\mathbb{Z}^d}$ with the corresponding eigenvalues $\{\mu_k := 1/v_{d,s}[k]^2\}_{k\in\mathbb{Z}^d}$. □

A.7 Proof of Lemma 5.1

Proof: Let us denote the Hilbert subspace of permutation invariant functions by $\widetilde{\mathcal{H}}^{perm}\subset\mathcal{H}$. We note that $\mathcal{K}^{perm}(.,t)\in\widetilde{\mathcal{H}}^{perm}$, since $\mathcal{K}^{perm}(\sigma.,t) = \frac{1}{d!}\sum_{\sigma'\in S_d}\mathcal{K}(\sigma\sigma'.,t) = \frac{1}{d!}\sum_{\sigma''\in S_d}\mathcal{K}(\sigma''.,t) = \mathcal{K}^{perm}(.,t)$, where we changed the summation variable to $\sigma'' := \sigma\sigma'$. Given $f\in\widetilde{\mathcal{H}}^{perm}$, it also follows that $f(t) = \frac{1}{d!}\sum_{\sigma\in S_d}f(\sigma t) = \left\langle f, \frac{1}{d!}\sum_{\sigma\in S_d}\mathcal{K}(.,\sigma t)\right\rangle_{\mathcal{H}} = \langle f,\mathcal{K}^{perm}(.,t)\rangle_{\mathcal{H}}$. This shows $\mathcal{K}^{perm}$ is the reproducing kernel for $\widetilde{\mathcal{H}}^{perm}$, therefore $\mathcal{H}^{perm} = \widetilde{\mathcal{H}}^{perm}$ by uniqueness. □

A.8 Proof of Theorem 5.2

Proof: For convenience, in the following we let $\mathcal{X} := [0,1]^d$ and $\mathcal{X}_0$ be the sorted part of $\mathcal{X}$. Take $f\in\mathcal{H}^{perm}$, then we have $f\in\mathcal{H}$ by Lemam 5.1. We consider the construction of interpolent $\hat{f}_n\in\mathcal{H}$ given the data $\{(s_i,y_i)\}_{i=1}^n$ where $s_i\in[0,1]^d$ are assumed to be sorted, and $y_i := f(s_i)$. Since s_i are sorted, we can also write $s_i := \operatorname{sort} t_i$ for some $t_i\in[0,1]^d$. Let $\mathcal{L}\subset\mathcal{H}$ be the subspace spanned by $\{\mathcal{K}(.,s_i)\}_{i=1}^n$; $\mathcal{L}$ is closed, so we have the orthogonal decomposition $\mathcal{H} = \mathcal{L}\oplus\mathcal{L}^\perp$. The orthogonal projection $P:\mathcal{H}\to\mathcal{L}$ is linear, and Pf is uniquely determined by the values $\{y_i\}_{i=1}^n$, so we can write $\hat{f}_n = Pf = \sum_{i,j=1}^n\alpha_i^j y_j\mathcal{K}(.,s_i)$ for some $\alpha_i^j\in\mathbb{R}$. Therefore,

$$\left|f(t)-\hat{f}_n(t)\right| = \left|\left\langle f, \mathcal{K}(.,t) - \sum_{i,j=1}^n\alpha_i^j\mathcal{K}(.,s_j)\mathcal{K}(t,s_i)\right\rangle_{\mathcal{H}}\right| \le \|f\|_{\mathcal{H}}\cdot\left\|\mathcal{K}(.,t) - \sum_{i,j=1}^n\alpha_i^j\mathcal{K}(.,s_j)\mathcal{K}(t,s_i)\right\|_{\mathcal{H}}\,.$$

Let $u_j^* := \sum_{i=1}^n \alpha_i^j \mathcal{K}(t, s_i)$ and define $Q_t : \mathbb{R}^n \to \mathbb{R}$:

$$Q_t(u) := \left\| \mathcal{K}(., t) - \sum_{j=1}^n u_j \mathcal{K}(., s_j) \right\|_{\mathcal{H}}^2 = \mathcal{K}(t,t) - 2\sum_{j=1}^n u_j \mathcal{K}(t, s_j) + \sum_{i,j=1}^n u_i u_j \mathcal{K}(s_i, s_j).$$

Then we can write the above inequality more compactly as

$$\left| f(t) - \hat{f}_n(t) \right| \le \|f\|_{\mathcal{H}} \cdot \sqrt{Q_t(u^*)}.$$

So far, $t \in \mathcal{X}$ is arbitrary, but we will focus on the case where $t \in \mathcal{X}_0$ in the following. It is also useful to define:

$$h_n := \sup_{t \in \mathcal{X}_0} \min_{i=1,\cdots,n} \|t - s_i\|_2.$$

Since $\mathcal{K}$ is positive semidefinite, Q_t is a convex function (although, not necessarily *strictly* convex), and any stationary points of Q_t is a global minimum. It follows that u^* is a global minimum, where

$$\frac{\partial Q_t}{\partial u_k}(u^*) = -2\mathcal{K}(t, s_k) + 2\sum_{j=1}^n u_j^* \mathcal{K}(s_j, s_k) = -2\mathcal{K}(t, s_k) + 2\sum_{i,j=1}^n \alpha_i^j \mathcal{K}(t, s_i)\mathcal{K}(s_j, s_k) = 0.$$

The last line follows from the fact that $P\mathcal{K}(., s_k) = \mathcal{K}(., s_k)$. From Theorem 3.14 of Wendland (2004), since $\mathcal{X}_0 \subset \mathbb{R}^d$ satisfies an (r_d, θ_d)–interior cone condition for some parameters $\theta_d \in (0, \pi/2), r_d > 0$ depending on the dimension d, at any $t \in \mathcal{X}_0$, we can find $\{\widetilde{u}_j(t)\}_{j=1}^n$ such that

- $\sum_{j=1}^n \widetilde{u}_j(t) p(s_j) = p(t)$, for all $p \in \pi_{2\nu}(\mathbb{R}^d)$,
- $\sum_{j=1}^n |\widetilde{u}_j(t)| \le B_1$,
- $\widetilde{u}_j(t) = 0$ provided that $\|t - s_j\|_2 > B_2 h_n$,

for some constants B_1, B_2 depending only on the dimension d (through the parameters r_d, θ_d). Here, $\pi_{2\nu}(\mathbb{R}^d)$ denotes the space of polynomials on $\mathbb{R}^d$ of degree at most 2ν. Because u^* is a global minimum, $Q_t(u^*) \le Q_t(\widetilde{u}(t))$, and we have

$$\begin{aligned}
Q_t(u^*) \le Q_t(\widetilde{u}) &= \mathcal{K}(t,t) - 2\sum_{j=1}^n \widetilde{u}_j(t)\mathcal{K}(t, s_j) + \sum_{i,j=1}^n \widetilde{u}_i(t)\widetilde{u}_j(t)\mathcal{K}(s_i, s_j) \\
&= \mathcal{K}(t,t) - 2\sum_{j=1}^n \widetilde{u}_j(t) \left(\sum_{|\alpha|<2\nu} \frac{(t-s_j)^\alpha}{\alpha!} \partial_2^\alpha \mathcal{K}(t,t) + \sum_{|\alpha|=2\nu} R_{0,\alpha}(t,t;t,s_j)(t-s_j)^\alpha \right) \\
&+ \sum_{i,j=1}^n \widetilde{u}_i(t)\widetilde{u}_j(t) \left(\sum_{|\alpha|+|\beta|<2\nu} \frac{(t-s_i)^\alpha (t-s_j)^\beta}{\alpha!\beta!} \partial_1^\alpha \partial_2^\beta \mathcal{K}(t,t) + \sum_{|\alpha|=2\nu} R_{\alpha,\beta}(t,t;s_i,s_j)(t-s_i)^\alpha (t-s_j)^\beta \right) \\
&= -2\sum_{j=1}^n \sum_{|\alpha|=2\nu} \widetilde{u}_j(t) R_{0,\alpha}(t,t;t,s_j)(t-s_j)^\alpha \\
&\quad + \sum_{i,j=1}^n \sum_{|\alpha|+|\beta|=2\nu} \widetilde{u}_i(t)\widetilde{u}_j(t) R_{\alpha,\beta}(t,t;s_i,s_j)(t-s_i)^\alpha (t-s_j)^\beta
\end{aligned}$$

where we use the fact that $\mathcal{K} \in C^{2\nu}(\mathcal{X} \times \mathcal{X})$ and apply multivariate Taylor's Theorem from any point $(t,t) \in \mathcal{X}_0 \times \mathcal{X}_0$ to (t, s_i) and (s_i, s_j), and the last line follows from the reproducing properties of $\widetilde{u}$ for $\pi_{2\nu}(\mathbb{R}^d)$ polynomials. Here, we denote the remainder by

$$R_{\alpha,\beta}(t_1, t_2; t_1', t_2') := \frac{1}{\alpha!\beta!} \partial_1^\alpha \partial_2^\beta \mathcal{K}(\xi_1, \xi_2)$$

for some $(\xi_1, \xi_2) \in \mathcal{X}_0 \times \mathcal{X}_0$ on the line connecting any $(t_1, t_2) \in \mathcal{X}_0 \times \mathcal{X}_0$ and $(t_1', t_2') \in \mathcal{X}_0 \times \mathcal{X}_0$. Note that

$$|R_{\alpha,\beta}(t_1, t_2; t_1', t_2')| \leq \max_{\xi_1,\xi_2 \in \mathcal{X}_0} \max_{|\alpha|+|\beta|=2\nu} \frac{1}{\alpha!\beta!} \left|\partial_1^\alpha \partial_2^\beta \mathcal{K}(\xi_1, \xi_2)\right| =: C_\mathcal{K}$$

which is finite because $\mathcal{X}_0$ is compact. From the vanishing property, $\widetilde{u}_j(t) = 0$ if $\|t - s_j\|_2 > B_2 h_n$, so we can bound $|t - s_i|^\alpha \leq \|t - s_i\|_2^{|\alpha|} \leq (B_2 h_n)^{|\alpha|}$. Continuing the above calculation with the triangle inequality yields

$$\begin{aligned} Q_t(u^*) &\leq 2 \cdot \binom{2\nu + d}{d} \cdot C_\mathcal{K} \, (B_2 h_n)^{2\nu} \sum_{j=1}^n |\widetilde{u}_j(t)| + \binom{2\nu + 2d}{2d} \cdot C_\mathcal{K} \, (B_2 h_n)^{2\nu} \left(\sum_{j=1}^n |\widetilde{u}_j(t)|\right)^2 \\ &\leq \left(\binom{2\nu + d}{d} + \binom{2\nu + 2d}{2d} B_1\right) B_1 B_2^{2\nu} C_\mathcal{K} h_n^{2\nu} \end{aligned}$$

where we have used the property that $\sum_{j=1}^n |\widetilde{u}_j(t)| \leq B_1$ to conclude the final line. Overall, we have

$$\left|f(t) - \hat{f}_n(t)\right| \leq B\|f\|_\mathcal{H} \cdot h_n^\nu, \quad \forall t \in \mathcal{X}_0$$

for some constant B depending only on the dimension d (through B_1, B_2) and the kernel (through $C_\mathcal{K}$). In fact, this bound is uniform. Since $f \in \mathcal{H}^{perm}$, $f \circ \mathrm{sort} = f$; also, $\hat{f}_n \circ \mathrm{sort} \in \mathcal{H}^{\mathrm{sort}}$ by definition. So we have

$$\left\|f - \hat{f}_n \circ \mathrm{sort}\right\|_{L^2(\mathcal{X},\mathbb{P})}^2 = \int_\mathcal{X} \left|f(\mathrm{sort}\, t) - \hat{f}_n(\mathrm{sort}\, t)\right|^2 dt \leq B\|f\|_\mathcal{H} \cdot h_n^\nu \to 0$$

as $n \to \infty$, for an appropriately chosen sequence of datasets. Therefore, there exists a sequence $\left\{\hat{f}_n \circ \mathrm{sort}\right\}_{n=1}^\infty \subset L^2(\mathcal{X}, \mathbb{P}) \cap \mathcal{H}^{\mathrm{sort}}$ converging to $f \in \mathcal{H}^{perm}$, which means $\mathcal{H}^{perm} \subset \overline{\mathcal{H}}^{\mathrm{sort}}$. □

A.9 Figures and Tables

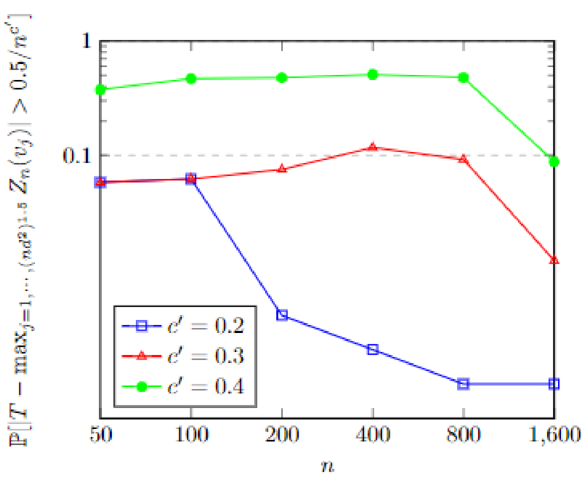

Figure 1: Convergence of $\max_{j=1,\cdots,(nd^2)^{1.5}} Z_n(v_j)$ to T for $d = 2$ using $\{v_j\}_{j=1,\cdots,(nd^2)^{1.5}}$ test points randomized i.i.d. from F.

	(μ_1, σ_1^2)	(0.5, 0.01)		(0.4, 0.01)		(0.5, 0.05)	
n	α	95%	99%	95%	99%	95%	99%
100	Cov	**0.96**	**0.99**	< 0.01	< 0.01	0.09	0.29
	Pow	0.04	0.01	$>$ **0.99**	$>$ **0.99**	**0.91**	**0.71**
200	Cov	**0.95**	**0.99**	< 0.01	< 0.01	< 0.01	0.01
	Pow	0.05	0.01	$>$ **0.99**	$>$ **0.99**	$>$ **0.99**	**0.99**
300	Cov	**0.95**	**0.99**	< 0.01	< 0.01	< 0.01	< 0.01
	Pow	0.05	0.01	$>$ **0.99**	$>$ **0.99**	$>$ **0.99**	$>$ **0.99**

Table 1: Simulation results with $d = 2$.

	μ_1	0.5		0.4	
d	α	95%	99%	95%	99%
3	Cov	**0.96**	**0.99**	< 0.01	0.01
	Pow	0.04	0.01	$>$ **0.99**	**0.99**
4	Cov	**0.95**	$>$ **0.99**	< 0.01	0.01
	Pow	0.05	< 0.01	$>$ **0.99**	**0.99**
5	Cov	**0.96**	**0.99**	0.03	0.11
	Pow	0.04	0.01	**0.97**	**0.89**

Table 2: Simulation result with $n = 100$.

t	$\hat{f}(t)$ bias	$\tilde{f}(t)$ bias	$\hat{f}(t)$ variance	$\tilde{f}(t)$ variance
	(all numbers below $\times 10^{-3}$)			
(0,0.5,1)	-0.1202	-0.3107	0.0523	0.0170
(0, 0.25, 0.5)	-1.0118	-0.9651	0.0802	0.0282
(0, 0.25, 0.5, 0.75)	-0.3838	-0.2504	0.0328	0.0082
(0, 0.25, 0.5, 0.75, 0.75)	-0.0913	-0.1394	0.0114	0.0026
(0, 0.25, 0.5, 0.75, 1)	-0.0332	-0.1423	0.0087	0.0017

Table 3: Bias and variance comparisons with $n = 10000$, $h = 3\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$ and 1000 replications.

d	Sorted	Invariant	Standard
	(all numbers below $\times 10^{-3}$)		
2	1.985	1.095	2.079
3	3.580	1.018	3.012
4	3.954	0.796	4.142
5	4.857	0.895	6.415

(a) KRR with Sobolev kernel $\mathcal{K}_{d,s}$

d	Sorted	Invariant	Standard
2	0.115	0.118	0.129
3	0.176	0.104	0.194
4	0.125	0.080	0.154
5	0.095	0.077	0.134

(b) KRR with 1–layer neural network kernel $\mathcal{K}_{NN}$

Table 4: Average SMSE comparison of sorted vs. permutation invaraint vs. standard KRR for an example of $f \in W^{s,2}_{per,perm}([0,1]^d)$ using the Sobolev kernel and 1–layer neural network kernel.

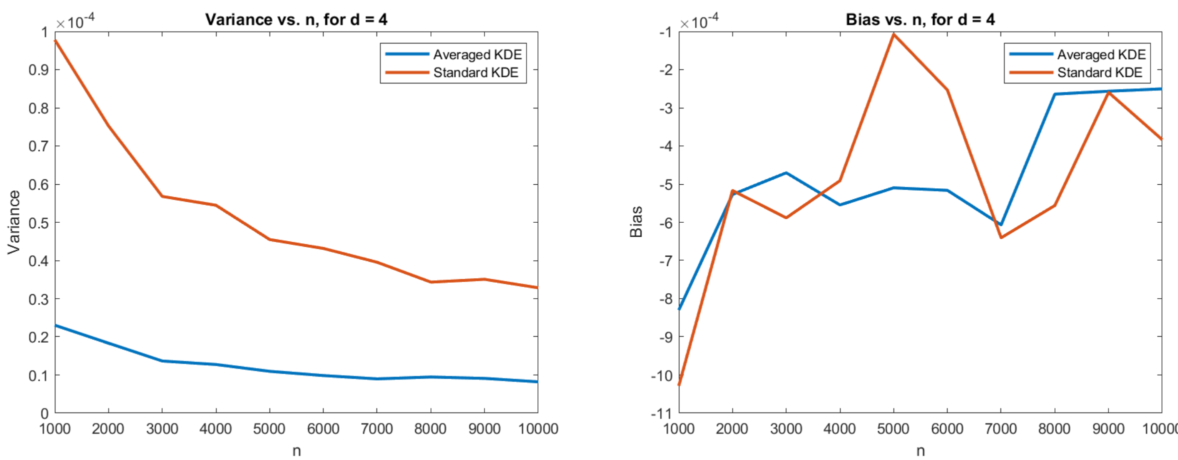

Figure 2: Variance and bias at $t = (0, 0.25, 0.5, 0.75)$ as n increases from 1000 to 10000 using the bandwidth of $h = 3\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$ and 1000 replications. The variance of the averaged kernel density estimator is consistently smaller than that of the standard kernel density estimator throughout.