Minimax-Optimal Kernel Two-sample Testing in Sub-quadratic Time

Ikjun Choi [1] **Shourya Pandey** [2] **Purnamrita Sarkar** [1]

Abstract

Kernel two-sample tests based on Maximum Mean Discrepancy (MMD) are widely used, but the standard quadratic time MMD statistic requires evaluating $\Theta(N^2)$ kernel pairs where N is the pooled sample size. Several subquadratic approximated MMD tests have been proposed to alleviate this issue, but the existing methods either compromise the test power or assume additional distributional constraints to achieve power comparable to the quadratic-time MMD test. Motivated by this gap, we propose a sub-quadratic time computational method to approximate the MMD test, achieving minimax rate-optimal power with minimal distributional assumptions. Building on ideas in fast kernel density estimation and exponentially convergent trapezoidal rule in numerical analysis, our approximation reduces all-pairs kernel summation to fast Fenwick-tree prefix-sum queries. For fixed dimension, this yields subquadratic-time algorithms for several common characteristic kernels, including Laplace, Matérn, Gaussian, and inverse multiquadric kernels. The approximation error is controlled tightly enough to have the equivalent power as the minimax-optimal quadratic-time unapproximated MMD test. The same implementation remains subquadratic for the dimension $d = o(\log N/\log\log N)$, and can be paired with dimensionality-reduction algorithms when the signal is low-dimensional.

1. Introduction

Kernel two-sample testing (Gretton et al., 2009, 2012) is a central nonparametric tool for deciding whether two samples come from the same distribution. Given samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_m \sim Q$, kernel tests based on the maximum mean discrepancy (MMD) compare the kernel mean embeddings of P and Q in a reproducing kernel Hilbert space. When the kernel is characteristic, the MMD is zero if and only if $P = Q$, making MMD tests broadly applicable without explicit density estimation. On top of its flexibility, then MMD test is known to be optimal against smooth alternatives from the perspective of minimax testing framework (Kim & Schrab, 2026; Li & Yuan, 2024; Schrab et al., 2023).

Despite these guarantees, the standard MMD statistic costs $\Theta(N^2 d)$ for $N = n+m$ samples in $\mathbb{R}^d$. Faster subsampling-based tests have been proposed (Gretton et al., 2012; Schrab et al., 2022; Yamada et al., 2019; Zaremba et al., 2013), but they either lose testing power or require quadratic time to achieve minimax optimality (Domingo-Enrich et al., 2023, Proposition 2). Other scalable methods exist (Chatalic et al., 2022, 2025; Chwialkowski et al., 2015; Kirchler et al., 2020), but their computation–power tradeoffs are less fully characterized. Random Fourier features (Choi & Kim, 2024) and kernel thinning (Domingo-Enrich et al., 2023; Dwivedi & Mackey, 2021) are notable exceptions, showing subquadratic minimax optimality for L_2- and MMD-separation, respectively, under favorable tail/kernel regimes. This leaves the basic question:

Can one retain the minimax-optimal power of the quadratic MMD test while computing the statistic in genuinely subquadratic time under minimal distributional assumptions?

We answer this question affirmatively for a broad class of commonly used characteristic kernels. Our starting point is the observation that MMD computation reduces to evaluating a small number of kernel lower-triangular sums. For kernels with *separable* structure, these sums can be computed without explicitly forming the kernel matrix. In one dimension, this viewpoint was used by (Bodenham & Kawahara, 2023), who reduce MMD computation to lower-triangular kernel sums. Using Fenwick trees (Fenwick, 1994) and their multidimensional variants (Mishra, 2013; Wang et al., 2021) our algorithm runs in $O((2\log N)^d N)$ time and memory for fixed d, improving substantially over the naive $O(N^2 d)$ computation when $N \gg d$.

To go beyond exactly separable kernels, we introduce an approximation framework based on almost separable kernels. The approximation step is based on the exponentially

[1]Department of Statistics and Data Sciences, University of Texas at Austin, TX, USA [2]Department of Computer Science, University of Texas at Austin, TX, USA. Correspondence to: Ikjun Choi <ikjunchoi@utexas.edu>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

convergent trapezoidal rule for analytic integrands on the real line, a classical principle in numerical approximation theory (Trefethen & Weideman, 2014). Consequently, for fixed d, products of Laplace, Matérn, Gaussian, and inverse multiquadric kernels admit fast MMD computation with only polylogarithmic overhead in N. The key statistical point is that the approximation error can be made small enough not to affect the testing rate.

We also extend current lower and upper bounds of the minimax separation rate for two-sample testing to allow d to grow with n. Interestingly, our analysis shows that when $d = \Omega(\log N)$, the rate does not go to zero. On the flip side, for $d = o(\log N/\log\log N)$ our Fenwick-tree based algorithm takes subquadratic time for a broad family of kernels. In this regime, the lower and upper bounds match up to polylogarithmic factors.

Our contributions can be summarized as follows:

1. We develop a *fast exact algorithm* for MMD computation with orthantwise separable kernels. For product Laplace kernels and related *separable* kernels (see Definition 3.1), the algorithm computes the same quadratic MMD statistic in $O((2\log N)^d N)$ time, rather than $O(N^2 d)$ time.

2. We introduce an approximation framework for characteristic kernels and prove that Gaussian and inverse multiquadric kernels admit *exponentially accurate* separable approximations, while half-integer Matérn and Laplace kernels are exactly finite separable sums.

3. We prove that the resulting approximated MMD permutation test is non-asymptotically level α and has the *same* uniform separation rate as the *quadratic-time MMD test*, up to a constant factor. In particular, for fixed dimension, our method achieves minimax-rate optimality in near-linear time for both L^2-separated alternatives over Sobolev ball and MMD-separated alternatives.

4. We also analyze the *growing-dimensional* regime. The same implementation remains subquadratic when $d = o(\log N/\log\log N)$, and we show how the method can be combined with Principal Component Analysis when the signal lies in a low-dimensional latent subspace.

Taken together, these results show that the quadratic cost of the complete MMD statistic is not intrinsic for many standard kernels. By exploiting kernel structure and classical numerical approximation, one can compute a statistically faithful approximation to the full MMD test in subquadratic time without reducing the test to a weaker statistic.

Section 2 reviews two-sample testing, MMD, and permutation tests. Section 3 develops the fast separable-kernel summation algorithm, Section 4 proves approximation and minimax testing guarantees, Section 5 treats dimensionality reduction, and Section 6 presents experiments. For space concerns we defer the related work section to the Appendix A.

2. Preliminaries

Notation. For $x \in \mathbb{R}^d$, x_ℓ is its ℓth coordinate and $x \preceq y$ means $x_\ell \le y_\ell$ for all ℓ. The coordinate-wise rank vector of Z_i is $r_i = (r_{i,1}, \dots, r_{i,d})$. We use $\|\cdot\|_2, \|\cdot\|_\infty$ for function norms and $\|\cdot\|$ for Euclidean/operator norms. The Sobolev ball is denoted by $\mathcal{S}_d^s(R)$. Finally, $a_n \lesssim b_n$ means $a_n \le C b_n$ for a constant independent of n, and $a_n \asymp b_n$ means $(a_n \lesssim b_n) \wedge (b_n \lesssim a_n)$.

2.1. Two sample testing

Consider two distributions P, Q on a domain $\mathbb{S}$, with independent samples $\mathbf{X} = \{X_i\}_{i=1}^n \sim P^{\otimes n}$ and $\mathbf{Y} = \{Y_j\}_{j=1}^m \sim Q^{\otimes m}$. Let $\mathcal{P}(\mathbb{S})$ denote the set of pairs of distributions on $\mathbb{S}$. For a distance $d(\cdot,\cdot)$ and optional regularity classes $\mathcal{C}_0, \mathcal{C}_1 \subseteq \mathcal{P}(\mathbb{S})$, define $\mathcal{P}_0 := \{(P,Q) \in \mathcal{C}_0 : P = Q\}, \mathcal{P}_1(\epsilon) := \{(P,Q) \in \mathcal{C}_1 : d(P,Q) > \epsilon\}$. The two-sample testing problem is the decision problem

$$H_0 : (P,Q) \in \mathcal{P}_0 \qquad \text{vs.} \qquad H_1 : (P,Q) \in \mathcal{P}_1(\epsilon). \quad (1)$$

A test is a function $\varphi : \mathbb{S}^n \times \mathbb{S}^m \to \{0,1\}$, where $\varphi(\mathbf{X},\mathbf{Y}) = 1$ means rejecting H_0. Its type-I error under $(P,Q) \in \mathcal{P}_0$ is $\mathbb{P}_{P^{\otimes n} \otimes Q^{\otimes m}}(\varphi = 1)$, and its type-II error under $(P,Q) \in \mathcal{P}_1(\epsilon)$ is $\mathbb{P}_{P^{\otimes n} \otimes Q^{\otimes m}}(\varphi = 0)$. Unless stated otherwise, we take $\mathbb{S} = \mathbb{R}^d$ and assume balanced samples: $n \le m \le Cn$ for a universal constant $C \ge 1$.

Minimax testing framework. Adopting the minimax testing framework pioneered by Ingster (1987), we characterize how φ performs across different scenarios by considering a uniform guarantee for those errors of φ over the entire classes $\mathcal{P}_0$ and $\mathcal{P}_1(\epsilon)$. For a given level $\alpha \in (0,1)$, we say φ is a *level-α* test if $\sup_{(P,Q)\in\mathcal{P}_0} \mathbb{P}_{P\otimes Q}[\varphi = 1] \le \alpha$, and for such non-asymptotic level-α test φ, we define the risk function as the worst-case type II error, i.e., $R(\varphi;\epsilon) = \sup_{(P,Q)\in\mathcal{P}_1(\epsilon)} \mathbb{P}_{P\otimes Q}[\varphi = 0]$. Then, we define the minimax risk as $R^*(\epsilon) = \inf_{\varphi\in\Phi(\alpha)} R(\varphi;\epsilon)$, where $\Phi(\alpha) = \{\varphi : \sup_{(P,Q)\in\mathcal{P}_0} \mathbb{P}_{P\otimes Q}[\varphi = 1] \le \alpha\}$ is the set of all level-α tests. For a given $\beta \in (0, 1-\alpha)$, we define the *uniform separation rate* (Baraud, 2002; Schrab et al., 2023) $\rho_n(\varphi;\beta,\mathcal{C}_0,\mathcal{C}_1)$ of φ and the *minimax separation rate* ϵ_n^*:

$$\rho_n(\varphi;\beta,\mathcal{C}_0,\mathcal{C}_1) = \inf\{\epsilon > 0 : R(\varphi;\epsilon) \le \beta\},$$
$$\epsilon_n^*(\beta,\mathcal{C}_0,\mathcal{C}_1) = \inf\{\epsilon > 0 : R^*(\epsilon) \le \beta\}.$$

If a test φ satisfies $\rho_n \le C\epsilon_n^*$ for some constant $C > 0$, we say φ is *minimax rate optimal*.

2.2. Maximum Mean Discrepancy

Given a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ equipped with a positive definite kernel k, the MMD dis-

tance (Gretton et al., 2012) between two distributions P and Q is defined as $\mathrm{MMD}(P,Q;k) := \|\mathbb{E}_{X\sim P}[k(X,\cdot)] - \mathbb{E}_{Y\sim Q}[k(Y,\cdot)]\|_{\mathcal{H}_k}$, as a measure of discrepancy between two distributions P and Q. If the kernel k is characteristic (Sriperumbudur et al., 2010), it is known that $\mathrm{MMD}(P,Q;k) = 0$ if and only if $P = Q$. An example of such a kernel is the Gaussian kernel: $k(x,y) = \frac{1}{\lambda\sqrt{\pi}}\exp(-|x-y|^2/\lambda^2)$ for some bandwidth $\lambda > 0$. This motivates the two sample test of the form $\Delta_n = \mathbf{1}\{T > q_n\}$, where T is an estimator of squared $\mathrm{MMD}(P,Q;k)$ and q_n is a threshold. For example, a standard unbiased estimator of $\mathrm{MMD}^2(P,Q;k)$ is defined as

$$\widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};k) = \frac{1}{n(n-1)}\sum_{i\neq j} k(X_i,X_j) + \frac{1}{m(m-1)}\sum_{i\neq j} k(Y_i,Y_j) - \frac{2}{nm}\sum_{i,j} k(X_i,Y_j).$$

Among two standard ways to choose the threshold q_n: (i) the asymptotic distribution of the estimator under the null hypothesis, or (ii) permutation method, we adopt the second approach here to enjoy non-asymptotic control of the type I error.

Permutation test. Given samples $\mathbf{X},\mathbf{Y}$, let $\mathbf{Z} = (Z_1,\dots,Z_{n+m}) = (X_1,\dots,X_n,Y_1,\dots,Y_m)$. For a permutation π, define $\mathbf{X}_\pi = (Z_{\pi(1)},\dots,Z_{\pi(n)})$, $\mathbf{Y}_\pi = (Z_{\pi(n+1)},\dots,Z_{\pi(n+m)})$, and $T_k^\pi = T_k(\mathbf{X}_\pi,\mathbf{Y}_\pi)$. Given B uniformly sampled permutations, the MMD permutation test is $\Delta_\alpha^B(\mathbf{X},\mathbf{Y};k) = \mathbf{1}\{T_k > q_{1-\alpha,B}\}$, where $q_{1-\alpha,B}$ is the $\lceil(1-\alpha)(B+1)\rceil$-th order statistic of $\{T_k, T_k^{\pi_1},\dots,T_k^{\pi_B}\}$. This test is exact level-α under label exchangeability (Hoeffding, 1952), and has essentially the same power as the full permutation test once $B \geq B^*$, where B^* is independent of n,d (Kim et al., 2022; Schrab et al., 2023).

3. Fast MMD Computation with Separable Kernel

A naive implementation of the MMD statistic requires $O(N^2d)$ time, which is prohibitive for large data. In this section, we present a fast algorithm for computing the MMD statistic equipped with a separable kernel, which reduces the time complexity to near linear time $O(N\log^{(d-1)\vee 1}N)$ when d is fixed. A first key observation is that the MMD statistic can be expressed as a kernel lower triangular sum (Bodenham & Kawahara, 2023). Note that the estimator above can be rewritten as

$$\widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};k) = \frac{2}{n(n-1)}K_{XX} + \frac{2}{m(m-1)}K_{YY} - \frac{2}{nm}K_{XY},$$

where $K_{XX} = \sum_{j<i} k(x_i,x_j)$, $K_{YY} = \sum_{j<i} k(y_i,y_j)$ and $K_{XY} = \sum_{i=1}^n\sum_{j=1}^m k(x_i,y_j)$. Note that the elements in K_{XX}, K_{YY}, K_{XY} form the lower triangular part of the kernel matrix of Z, i.e., $\{k(z_i,z_j) : j < i\}$. Indeed, note that the lower triangular part of the kernel matrix $K_{ZZ} = \sum_{j<i} k(z_i,z_j)$ can be written as $K_{ZZ} = K_{XX} + K_{YY} + K_{XY}$. Therefore, we can compute the estimator as $\widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};k) = \frac{2}{n(n-1)}K_{XX} + \frac{2}{m(m-1)}K_{YY} - \frac{2}{nm}(K_{ZZ} - K_{XX} - K_{YY})$. Observe that all K_{XX}, K_{YY}, K_{ZZ} are kernel lower triangular sums, so the computation of the MMD statistic can be reduced to computing kernel lower triangular sums (Bodenham & Kawahara, 2023). In what follows, we will focus on computing K_{ZZ} for proper kernel, and the same algorithm applies to compute K_{XX} and K_{YY}.

3.1. Kernel matrix computation with separable kernel

In this section, we describe how to compute the kernel lower triangular sum efficiently using the Fenwick data structure (Fenwick, 1994), which is a powerful tool for computing prefix sums and updating elements in logarithmic time. We start by a pedagogical example for the case of $d = 1$, and then provide the full algorithm for arbitrary d. We first define a family of kernels which is compatible with the Fenwick data structure.

Definition 3.1 (Piecewise Separable kernel, Orthantwise Separable kernel)**.** For a one-dimensional kernel $k : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, we say k is piecewise separable (PS) if it can be expressed as

$$k(x,y) = \mathbf{1}\{y \leq x\}f_-(x)g_-(y) + \mathbf{1}\{y > x\}f_+(x)g_+(y),$$

where f_+, f_-, g_+, g_- are some functions defined on $\mathbb{R}$. More generally, we call a kernel $k : \mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ orthantwise separable (OS) if it can be expressed as

$$k(x,y) = \sum_{s\in\{+,-\}^d} \mathbf{1}\{y \in Q_s(x)\}f_s(x)g_s(y), \quad (2)$$

where $Q_s(x) = \{y\in\mathbb{R}^d : \mathrm{sign}(y_i - x_i)^1 = s_i, \forall i\}$ indicate which orthant y is in with respect to x, and f_s, g_s are some functions defined on $\mathbb{R}^d$ for each $s\in\{+,-\}^d$.

For example, the one-dimensional Laplace kernel $k(x,y) = \exp(-|x-y|/\sigma)$ is PS with $f_+ = g_- = \exp(\cdot/\sigma)$ and $f_- = g_+ = \exp(-\cdot/\sigma)$. Since products of PS kernels are OS, the product Laplace kernel $k(x,y) = \prod_{i=1}^d \exp(-|x_i - y_i|/\sigma_i)$ is OS. For such kernels, we compute the lower-triangular kernel sum $K_{ZZ} = \sum_{1\leq j<i\leq N} k(z_i,z_j) = \sum_{i=1}^N C_i$, where $C_i = \sum_{j<i} k(z_i,z_j)$, by iterating over rows. The key observation

[1] We define $\mathrm{sign}(t) = +$ if $t > 0$, and $\mathrm{sign}(t) = -$ if $t \leq 0$ for $t\in\mathbb{R}$.

is that contributions from points in each orthant reduce to prefix sums, which Fenwick trees can update and query in logarithmic time. In one dimension, if $z_1 \le z_2 \le z_3 \le z_4$, then instead of computing $C_4 = \sum_{j=1}^{3} k(z_4, z_j)$ directly, we use $C_4 = f_-(z_4) \sum_{j=1}^{3} g_-(z_j)$, which is efficient once the prefix sum $\sum_{j=1}^{3} g_-(z_j)$ is stored.

3.2. Fast algorithm with Fenwick structure

For $d = 1$, after computing ranks r_i in $O(N \log N)$ time, the i-th row sum is

$$C_i = f_-(z_i) \sum_{j<i} \mathbf{1}\{z_j \le z_i\} g_-(z_j) + f_+(z_i) \sum_{j<i} \mathbf{1}\{z_j > z_i\} g_+(z_j).$$

Thus, rather than summing over all $j < i$, it suffices to query prefix and suffix sums of previously inserted values. A Fenwick tree supports such prefix queries and point updates in $O(\log N)$ time. Maintaining two trees for g_- and g_+, we compute the two sums above, add C_i, and then insert $g_-(z_i)$ and $g_+(z_i)$ at rank r_i. Correctness follows from the invariant that, after $i-1$ updates, the tree stores exactly the contributions of the first $i-1$ points.

A generalization of the above algorithm is straightforward for the multi-dimensional case. Suppose that there exist a multidimensional array structure $\texttt{T} = (\texttt{T[k]})_{k \in \{1,\dots,N\}^d}$ which can be updated and queried as follows: for $a = (a_1, \dots, a_d)$, we define query for a-th box prefix sum as $\texttt{T.query}(a) := \sum_{k_1=1}^{a_1} \cdots \sum_{k_d=1}^{a_d} \texttt{T}[k_1, \dots, k_d]$, and the update operation is defined as $\texttt{T.update}(a, v)$ which updates the a-th element of $\texttt{T}$ as $\texttt{T[a]} + v$. Note that at step i, the algorithm satisfies $\texttt{T.query}(r_i) = \sum_{j<i} \mathbf{1}\{r_{j,1} \le r_{i,1}, \dots, r_{j,d} \le r_{i,d}\} g(z_j)$, where $r_{i,k}, r_{j,k}$ are the coordinate-wise rank of z_i, z_j in the k-th dimension, respectively.

The direct implementation has time $O((4 \log N)^d N)$ and memory $O(N^d)$. Appendix D gives the pseudocode of this algorithm and the Fenwick data structure details; using symmetry, rank flips, and compressed trees improves this to $O\big((2 \log N)^d N\big)$ time and memory.

3.3. Approximation with almost separable kernels

In this section, we expand the class of kernels compatible with our algorithm. We consider kernels that can be approximated by a sum of PS or OS kernels sacrificing a small approximation error.

Definition 3.2 (Almost Piecewise Separable kernel, Almost Orthantwise Separable kernel)**.** For a one-dimensional kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we say k is (R, δ)-almost piecewise separable ((R, δ)-APS) if it can be approximated by a sum of R piecewise separable kernels $\tilde{k} = \sum_{r=1}^{R} \tilde{k}_r$ such that $\sup_{x,y} |k(x, y) - \tilde{k}(x, y)| \le \delta$, for some $R \in \mathbb{N}$ and $\delta \ge 0$. Similarly, we say a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is (R, δ)-almost orthantwise separable ((R, δ)-AOS)[2] if it can be approximated by a sum of R orthantwise separable kernels $\tilde{k} = \sum_{r=1}^{R} \tilde{k}_r$ such that $\sup_{x,y} |k(x, y) - \tilde{k}(x, y)| \le \delta$.

If k is (R, δ)-APS/AOS with approximation $\tilde{k} = \sum_{r=1}^{R} \tilde{k}_r$, then $\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}; \tilde{k})$ can be computed in $O(R\, 2^d N \log^{(d-1)\vee 1} N)$ time, since $\sum_{j<i} \tilde{k}(z_i, z_j) = \sum_{r=1}^{R} \sum_{j<i} \tilde{k}_r(z_i, z_j)$ reduces to R OS lower-triangular sums. Thus R controls computational cost, while δ controls approximation quality; the next section studies this trade-off for common kernels.

3.4. Quality of approximation for characteristic kernels

We focus on characteristic kernels, for which $\mathrm{MMD}(P, Q) = 0$ iff $P = Q$, a key property for two-sample testing. Table 1 summarizes our (R, δ)-APS results for one-dimensional characteristic kernels, with details in Lemma C.1. For multidimensional data, we use product kernels $k(x, y) = \prod_{i=1}^{d} k_i(x_i, y_i)$, which remain characteristic when the factors are characteristic (Szabó & Sriperumbudur, 2018, Theorem 4). The next theorem bounds the AOS approximation quality of such product kernels from the APS approximations of the one-dimensional factors, allowing dimension-dependent bandwidths $\lambda_i > 0$ and non-identical k_i's. The proof is in Appendix C.1.

Theorem 3.3 (AOS kernel approximation on the product of characteristic kernels)**.** *Given bandwidth $\lambda = (\lambda_1, \dots, \lambda_d) \in (0, \infty)^d$, consider a product kernel scaled with bandwidth λ: $k_\lambda(x, y) = \prod_{i=1}^{d} (1/\lambda_i) k_i(x_i/\lambda_i, y_i/\lambda_i)$, where each k_i is a one-dimensional characteristic (R_i, δ_i)-APS kernel. Suppose that each k_i is bounded by $U_i := \sup_{x,y \in \mathbb{R}} |k_i(x, y)| < \infty$. Suppose that there exists a fixed constant $\delta_0 > 0$ such that $\delta_i \le \delta_0$ for all $i = 1, \dots, d$. Then for any (R_i, δ_i)-APS kernel k_i, the kernel k_λ is (R, δ)-AOS kernel with*

$$R = \prod_{i=1}^{d} R_i, \quad \delta = \frac{(U + \delta_0)^{d-1} d}{\lambda_1 \cdots \lambda_d} \delta_{\max},$$

where $U = \max_{i=1,\dots,d} U_i$ and $\delta_{\max} := \max_{i=1,\dots,d} \delta_i$.

Table 1 lists the most common characteristic kernels in the literature (Domingo-Enrich et al., 2023; Gorham & Mackey, 2017; Li & Yuan, 2024; Muandet et al., 2017; Schrab et al., 2023; Simon-Gabriel & Schölkopf, 2018). We show that they are PS or APS kernels with exponential decay of the

[2] (R, δ) characterization need not be unique. One can easily see that if k is (R, δ)-APS/AOS, it is also (R', δ')-APS/AOS for any $R' \ge R$ and $\delta' \ge \delta$.

Kernel	Shape of $k(x,y),\ x,y \in \mathbb{R}$	(R,δ)-APS approximation
Laplace	$\frac{1}{2}e^{-\lvert x-y\rvert}$	$(1,0)$
Matérn ($\nu = p+\frac{1}{2}$)	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\lvert x-y\rvert\right)^{\nu} K_\nu\left(\sqrt{2\nu}\lvert x-y\rvert\right)$	$(p+1,0)$
Gaussian	$\frac{1}{\sqrt{\pi}}e^{-\lvert x-y\rvert^2}$	$\left(\lceil 5\log(1/\delta)\rceil,\ \delta\right)^{\dagger}$
Inverse multiquadric	$(1+\lvert x-y\rvert^2/c^2)^{-\gamma}$	$\left(\lceil C_\gamma \log^3(1/\delta)\rceil,\ \delta\right)^{\dagger}$

Table 1. (R,δ)-APS approximations of characteristic kernels. $p \in \mathbb{Z}_{\geq 0}$, K_ν denotes the modified Bessel function of the second kind, and Γ is the Gamma function. Constants $c > 0$ and $\gamma > 0$ are fixed, and C_γ depends only on γ. The $(R,\delta)^{\dagger}$ approximation assumes $\delta < 0.1$.

approximation error $\delta < 0.1$. Appendix C.2 elaborates with precise lemma statement with arbitrary δ in terms of R, which allows us to achieve a good approximation with a small R.

4. Optimality of the Approximated MMD Test

In this section, we discuss the performance of the approximated MMD test, and show that the approximated MMD test achieves minimax rate optimality under various setting. We first quantify the level and the power of the approximated MMD test in the following lemma, which illustrates that the approximated MMD test has the same level α and almost equal power as the original MMD test if the approximation error δ decays sufficiently fast. We provide the proof of this lemma in Appendix C.4.

Lemma 4.1 (Test error of the approximated MMD test)**.** *Suppose that the characteristic kernel k_λ is a (R,δ)-APS or (R,δ)-AOS kernel with approximation $\tilde{k}_\lambda$, given bandwidth $\lambda \in (0,\infty)^d$. For $B \in \mathbb{N}$ and $\alpha \in (0,1)$, define a MMD permutation test $\Delta_\alpha^{B,\lambda}(\mathbf{X},\mathbf{Y};k) = \mathbf{1}\{T_{k_\lambda}(\mathbf{X},\mathbf{Y}) > q_{1-\alpha,B}(k)\}$ where $T_{k_\lambda}(\mathbf{X},\mathbf{Y}) := \widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};k_\lambda)$, and define an approximated test $\tilde{\Delta}_\alpha^{B,\lambda}(\mathbf{X},\mathbf{Y};k) := \Delta_\alpha^{B,\lambda}(\mathbf{X},\mathbf{Y};\tilde{k})$ by replacing k with $\tilde{k}$. Then, the test $\Delta_\alpha^{B,\lambda}(\mathbf{X},\mathbf{Y};\tilde{k})$ is non-asymptotic level-α test, i.e., $\mathbb{P}_{H_0}\left(\Delta_\alpha^{B,\lambda}(\mathbf{X},\mathbf{Y};\tilde{k}) = 1\right) \leq \alpha$ for any n and B. In addition, the power of the approximated test satisfies*

$$\begin{aligned}
&\mathbb{P}_{H_1}\Big(T_{k_\lambda}(\mathbf{X},\mathbf{Y}) - 8\delta > q_{1-\alpha,B}(k)\Big)\\
&\leq \mathbb{P}_{H_1}\Big(\Delta_\alpha^{B,\lambda}(\mathbf{X},\mathbf{Y};\tilde{k}) = 1\Big)\\
&\leq \mathbb{P}_{H_1}\Big(T_{k_\lambda}(\mathbf{X},\mathbf{Y}) + 8\delta > q_{1-\alpha,B}(k)\Big).
\end{aligned}$$

4.1. Classical regime: fixed dimension

We first consider the fixed-dimensional setting, where d does not grow with n. We show that the approximated MMD test with any product kernel from Lemma C.1, and with a suitable bandwidth, achieves minimax-rate optimality in near-linear time. For L_2-separation, assume P and Q have densities p, q satisfying $\mathcal{C}_1 = \{(P,Q) : p - q \in \mathcal{S}_d^s(\mathcal{R}),\ \|p\|_\infty, \|q\|_\infty \leq M\}$, where $M < \infty$ and

$$\mathcal{S}_d^s(\mathcal{R}) = \Big\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\hat{f}(\xi)|^2 d\xi \leq (2\pi)^d \mathcal{R}^2 \Big\}. \tag{3}$$

Here $\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-ix^\top \xi} dx$ and $s > 0$ is the smoothness parameter.[3] As shown in Li & Yuan (2024); Schrab et al. (2023), the minimax separation rate is $\epsilon_n^* \asymp n^{-2s/(4s+d)}$. The next theorem shows that the approximated MMD test attains this rate up to constants, provided $\delta \lesssim n^{-4s/(4s+d)}$, while remaining near-linear time. The proof is in Appendix C.5.

Theorem 4.2. *Consider the two sample testing problem defined in Equation (1) with $\mathcal{C}_1$ defined as above. Assume that $\alpha \in (0, e^{-1})$ and $\beta \in (0, 1-\alpha)$, and $B \geq 3\alpha^{-2}(\log(8/\beta) + \alpha(1-\alpha))$. Consider the product kernel k_λ defined in Theorem 3.3 with each k_i being Laplace, Matérn, Gaussian, or inverse multiquadric kernel with $\gamma > 1/2$ in Lemma C.1. With the choice of $\lambda_i^* = N^{-2/(4s+d)}$ for $i = 1, \ldots, d$, there exists a $\tilde{k}_{\lambda^*}$ which is (R,δ)-AOS approximation of k_{λ^*} with*

$$R = (p+1)^{m_k} \log^{3h_k + g_k} N, \quad \delta \leq C' N^{-4s/(4s+d)},$$

where $C' > 0$ is a universal constant, and l_k, m_k, g_k, h_k denote the number of Laplace, Matérn, Gaussian, and inverse multiquadric kernels producted in the kernel k_λ, respectively. The corresponding approximated test $\tilde{\Delta}_\alpha^{B,\lambda^*}$ with such $\tilde{k}_{\lambda^*}$ is minimax rate optimal, i.e.,

$$\rho_n(\tilde{\Delta}_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1) \leq C(\alpha, \beta, d, s, M, \mathcal{R}) \epsilon_n^*(\beta, \mathcal{C}_1).$$

Remark 4.3. A similar near-linear-time minimax rate optimality guarantee also holds with respect to the MMD metric. This is closely related to the result of Domingo-Enrich et al. (2023), which establishes minimax optimality for more general characteristic kernels under suitable tail decay assumptions. In our setting, we focus on products of the characteristic kernels listed in Table 1 and obtain a

[3] A smoothness condition such as a Sobolev, Hölder, or Nikol'skii constraint is necessary for nontrivial L_2 minimax testing; without it, the minimax separation rate remains trivial.

near-linear-time guarantee for this class without imposing such tail decay assumptions. We display the details of this claim and its proof in Appendix C.5.1 below.

4.2. Growing dimension regime

We generalize the problem to the setting where the dimension of the data can grow with n. Given growing sequence $d_n \to \infty$, we observe the sequence of two sets of i.i.d. samples in triangular array form, $\mathbf{X}^{(n)} := \{X_i^{(n)}\}_{i=1}^n$ and $\mathbf{Y}^{(n)} := \{Y_j^{(n)}\}_{j=1}^m$ drawn from $(P^{(n)})^{\otimes n}$ and $(Q^{(n)})^{\otimes m}$ defined on $\mathbb{R}^{d_n}$, respectively. We define $\mathcal{P}(\mathbb{R}^{d_n})$, $\mathcal{P}_0^{(n)}$ and $\mathcal{P}_1^{(n)}(\epsilon_n)$ with constraint $\mathcal{C}_1^{(n)}$ and corresponding testing problem in the similar manner to Section 2.1. We study the minimax separation rate ϵ_n^* as a function of n and the growing dimension d_n, and ask whether it is achievable in subquadratic time. To compare function classes across dimensions, we scale the constraints as $M_d = M^d$ and $\mathcal{R}_d = \mathcal{R}d^{s/2}$, reflecting the exponential growth of density bounds and the $d^{s/2}$ Sobolev scaling of the standard Gaussian. We then derive lower and upper bounds on ϵ_n^*, tight up to logarithmic factors, for MMD tests with product Gaussian or Laplace kernels. We provide the proof in Appendix C.6

Lemma 4.4. *Consider the two sample testing problem defined above, i.e. testing between $\mathcal{P}_0^{(n)}$ and $\mathcal{P}_1^{(n)}(\epsilon_n)$ with constraint $\mathcal{C}_1^{(n)} = \{(P,Q) : p - q \in \mathcal{S}_{d_n}^s(R_d),\ \|p\|_\infty, \|q\|_\infty \le M^d\}$, where $s \ge 1$, $\mathcal{R} > 1, M > 4$. For $\alpha, \beta, \eta \in (0,1)$ such that $\alpha + \beta + \eta < 1$, there exists a constant $C(\alpha, \beta, \eta, s, M) > 0$ such that the minimax separation rate ${\epsilon_n^{(n)}}^*$ satisfies*

$$ {\epsilon_n^{(n)}}^*(\beta, \mathcal{C}_1^{(n)}) \ge C(\alpha, \beta, \eta, s, M)(M^s\mathcal{R})^{\frac{d}{4s+d}} N^{-\frac{2s}{4s+d}} \wedge c. $$

Also, with the choice of bandwidth $\lambda_i^\star = \left(M^{d/2}\mathcal{R}^{-2d}d^{-s}N^{-1}\right)^{2/(4s+d)}$, there exists a constant $C(\alpha, \beta, s) > 0$ such that the uniform separation rate $\rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)})$ of the MMD test equipped with the product Gaussian kernel or the product Laplace kernel is upper bounded by

$$ \rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)}) \le C(\alpha, \beta, s)\left((M^s\mathcal{R})^{\frac{d}{4s+d}} N^{-\frac{2s}{4s+d}} + M^{d/2}N^{-1/2}\right). $$

We observe that where both lower bound and upper bound becomes nontrivial, i.e. ${\epsilon_n^{(n)}}^*, \rho_n \to 0$, is the regime $d = o(\log N)$. In this case, the MMD test with the product Gaussian kernel or the product Laplace kernel is minimax rate optimal up to a logarithmic factor of n. We note that the product Laplace kernel is OS and the approximated MMD test with the product Laplace kernel can be implemented within $O(N \log^d N)$ time without loss of power. Since this becomes subquadratic when $d = o(\log N/\log\log N)$, we proved the existence of a subquadratic minimax optimal test except the regime between $d = \Omega(\log N/\log\log N) \wedge d = o(\log N)$.

5. Latent Low-Dimensional Signal Model

We consider the following latent low-dimensional signal model. Let $\mathbf{M} \in \mathbb{R}^{d\times r}$, $\mathbf{M}^\top\mathbf{M} = \mathbf{I}_r$ and $r < d$, so that the columns of $\mathbf{M}$ form an orthonormal basis for a r-dimensional signal subspace of $\mathbb{R}^d$. We observe two d-dimensional random vectors X and Y of the form

$$ X = \mathbf{M}U + \xi, \qquad Y = \mathbf{M}V + \xi', \tag{4} $$

where $U, V \in \mathbb{R}^r$ are arbitrary latent random vectors, and $\xi, \xi' \in \mathbb{R}^d$ are mean zero noise random vectors. We assume that U, V, ξ, and ξ' are all independent. In this model, the difference between the laws of X and Y is entirely driven by the low-dimensional latent variables U and V through the subspace spanned by $\mathbf{M}$, while the ambient noise is shared across the two populations.

Remark 5.1. If the populations had unrelated bases $\mathbf{M}_X$ and $\mathbf{M}_Y$, then the latent coordinates would only be defined up to independent rotations; for example, $\mathbf{M}_X U = (\mathbf{M}_X\mathbf{O})(\mathbf{O}^\top U)$ for any orthonormal $\mathbf{O} \in \mathbb{R}^{k\times k}$. Thus differences between the observed laws could reflect either differences between the latent variables or a misalignment of the two subspaces. We therefore set $\mathbf{M}_X = \mathbf{M}_Y = \mathbf{M}$ and test for differences in the latent laws.

Algorithm 8. We handle the latent low-dimensional setting by first estimating the shared signal subspace via sample-splitting and doing Principal Component Analysis (PCA) on one half, and then applying the fast MMD permutation test to the held out data projected on the principal components (for detailed pseudocode and proofs see Appendix E).

Assumption 5.2 (Strong regularity)**.** Let $U \in \mathbb{R}^r$ have mean $\mu = \mathbb{E}U$ and covariance $\mathbf{\Sigma}_U$. There exist constants $C_{\mathrm{SR}} > 0$ and $\eta > 0$ such that, for every orthogonal projection $\mathbf{\Pi}$ and every $t > C_{\mathrm{SR}}\,\mathrm{rank}(\mathbf{\Pi})$, $\Pr\left(\|\mathbf{\Pi}\mathbf{\Sigma}^{-1/2}(U-\mu)\|_2^2 > t\right) \le C_{\mathrm{SR}}t^{-1-\eta}$.

Lemma 5.3. *Let ξ, ξ' both have the isotropic standard Gaussian distribution. Assume that X and Y satisfy the strong regularity assumption in Assumption 5.2 with $\eta > 0$. Let p_M, q_M be the densities of $\mathbf{M}^\top X$ and $\mathbf{M}^\top Y$. Assume $U - V \in \mathcal{S}_r^s(\mathcal{R})$ for some $s \ge 1$ with $\eta > 4s/k$, and $\|p_M - q_M\|_2 \gtrsim C_k N^{-2s/(4s+k)}$. Let $\lambda := \max(\|\mathbf{\Sigma}_U\|, \|\mathbf{\Sigma}_V\|)$ and $\gamma := \min(\lambda_r(\mathbf{\Sigma}_U), \lambda_r(\mathbf{\Sigma}_V))$ where $\mathbf{\Sigma}_U := \mathbb{E}[UU^T]$ and $\mathbf{\Sigma}_V = \mathbb{E}[VV^T]$. Then the L_2 distance of the projected densities using the projection matrix learned via PCA with Algorithm 8 reduces the original L_2 distance by at most a constant fraction as long as $\eta \ge 4s/r$ and $n \gtrsim C(r, \eta, C_{\mathrm{cov}}, \lambda/\gamma, m_r)d^{\frac{\eta(4s+r)}{\eta r - 4s}}$.*

Remark 5.4 (Interpreting the strong regularity condition). Assumption 5.2 is a polynomial-tail condition on whitened projections of Z, which is weaker than sub-gaussianity. In particular, taking $\mathbf{\Pi}$ to be rank one shows that every one-dimensional whitened projection has moments of order $p < 2(1+\eta)$, i.e. $\eta = 1$ corresponds roughly to existence of four moments and larger η to lighter tails. In the latent model (Equation 4) if ξ is isotropic Gaussian, and U satisfies Assumption 5.2, then x also satisfies Assumption 5.2 up to a change of constants.

Remark 5.5 (A concrete interpretation of Lemma 5.3). Take $s = 1$, $r = 10$, and $\eta = 2$. Then $\eta > 4s/r$, and Assumption 5.2 gives moments of all orders $p < 6$ for whitened projected norms. Lemma 5.3 implies that PCA is accurate at the testing scale when $n = \Omega(d^{7/4})$. Thus, for a ten-dimensional latent signal with moderately heavy-tailed observations, polynomial sample size in d suffices for the projected MMD test to behave as if the true subspace were known.

Although we take ξ, ξ' to be standard Gaussian, the results extend to different noise laws under mild regularity conditions; see Appendix E.

6. Experiments

We compare our approximation method with the complete MMD test and standard subquadratic baselines: linear MMD (Gretton et al., 2012), block MMD (Zaremba et al., 2013), incomplete U-statistics (Schrab et al., 2022; Yamada et al., 2019). Additionally, we adapt the CDQ kernel computation proposed by Langrené & Warin (2021) to the AOS and APS kernels and include it in our implementation. In each scenario, we run permutation two-sample tests with both Laplace and Gaussian kernels, varying $n = m$ from 500 to 4000. Each test uses $B = 100$ permutations and level $\alpha = 0.05$; empirical power is averaged over 100 trials. We use the median heuristic for bandwidth selection (Gretton et al., 2012) and report the time to compute a single MMD statistic. The Laplace kernel is PS, whereas the Gaussian kernel is approximated by an APS kernel with $R_i = 3$ per dimension, using the construction in Appendix C.3.2. Block MMD uses block size $n^{1/2}$, and incomplete U-statistics use $R = 100$ superdiagonals. The time and memory complexities of each method are summarized in Table 2. All experiments were run on an Apple MacBook Pro M3 Max with a 30-core GPU and 36GB RAM.

Scenarios. We first take a simple **gaussian mean shift** (GMS) alternative, where we generate $X_i \sim N(0,1)$ and $Y_i \sim N(0.2,1)$ for $i \in [n]$. We also consider a more complex alternative **perturbed uniform** (PU), mimicking the lower bound construction for Sobolev alternatives (see Appendix B). Specifically, we consider the alternative density $f_\theta(u) = \mathbf{1}_{[0,1]^d}(u) + c_d P^{-s} \sum_{\nu \in \{1,\dots,P\}^d} \theta_\nu \prod_{i=1}^d G(Pu_i - \nu_i)$, where θ_ν are fixed $\{1,-1\}^{P^d}$, and G is a smooth bump function defined as $G(u) = \exp(-1/(1-(4u+3))^2)\mathbf{1}_{(-1,-1/2)}(u) - \exp(-1/(1-(4u+1)^2)\mathbf{1}_{(-1/2,0)}(u)$. We use the parameter $P = 2$ and $s = 0.3$ for this experiment, and c_d is automatically chosen to make f_θ a nonnegative density. For the gaussian mean shift we use $d = 1$ and for the perturbed uniform $d = 2$. We set 100 permutations and 100 iteration for estimating the power.

Method	Test time	Memory
Full	$O(N^2 d + BN^2)$	$O(N^2)$
Linear	$O(BNd)$	$O(N)$
Block	$O(BNbd)$	$O(Nb)$
Incomplete	$O(BRNd)$	$O(NR)$
Fenwick	$O\big(B(2\log N)^d N\big)$	$O\big((2\log N)^d N\big)$
CDQ	$O\big(B(2\log N)^d N\big)$	$O\big((2\log N)^d N\big)$

Table 2. Complexity of permutation MMD-based tests. Here, $N = n + m$ is the total sample size, d is dimension, B is the number of permutations, R is the number of superdiagonals for incomplete U-statistics, b is the block size for block MMD.

Figure 1 compares six MMD-based two-sample testing methods. The top (bottom) row contains power (time) in Y axis across number of samples N on X axis. The left (right) half uses the Laplace (Gaussian) kernel, computed with our Fenwick tree (trapezoidal rule). The two families are GMS and PU. The figure shows that Fenwick retains power close to Full MMD while achieving much lower runtime, particularly for the Laplace kernel. CDQ also tends to have high power and good runtime, however its significantly slower than Fenwick when $d = 2$ showing a worse dependence on d. Incomplete U-statistics and Block MMD are faster than Full MMD, but their power is generally lower, especially in the perturbed uniform setting. Linear MMD is the fastest or among the fastest methods, but it has the weakest power across most settings.

PCA-MMD. For $N \in [2000, \dots 8000]$, we generate balanced samples $n = m = N/2$ in ambient dimension $d = 100$. The signal dimension is $k = 2$. In each trial, we set $\mathbf{M}$ to be a random orthonormal matrix $M \in \mathbb{R}^{d\times k}$ and generate from Eq 4. U, V come from the GMS or PU families. The ξ, ξ' are isotropic Gaussians with $\sigma = 0.1$ for GMS and $\sigma = 0.05$.

In the experiments, we estimate the top k principal components from the pooled sample $X \cup Y$, project both samples onto these components, and run the MMD permutation test on the projected k-dimensional data. We use $B = 100$ permutations and significance level $\alpha = 0.05$. The sample-splitting version in Algorithm 8 is used for the theoretical analysis; in the reported simulations, PCA is performed on

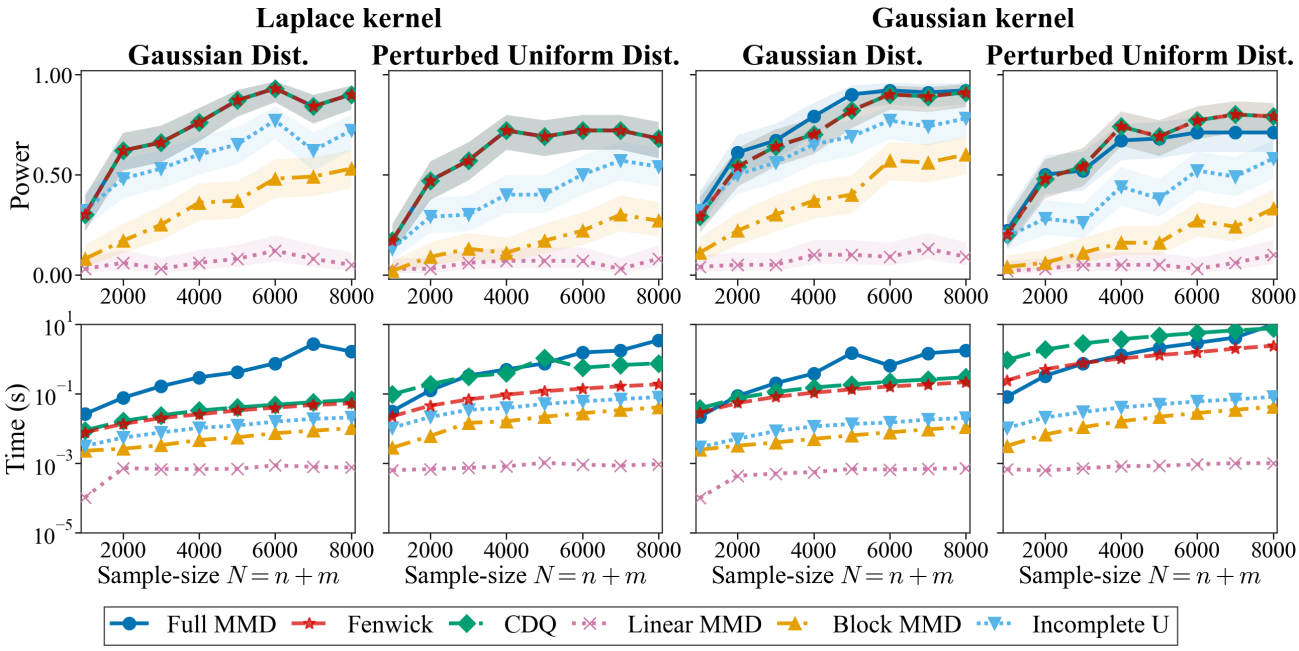

Figure 1. MMD permutation two sample testing experiments. The left (right) half uses the Laplace (Gaussian) kernel. The powers are averaged over 100 iterations, and the permutation test is run with $B = 100$ permutations. In the two upper-left plots, the Full MMD, Fenwick, and CDQ curves are identical and therefore overlap completely.

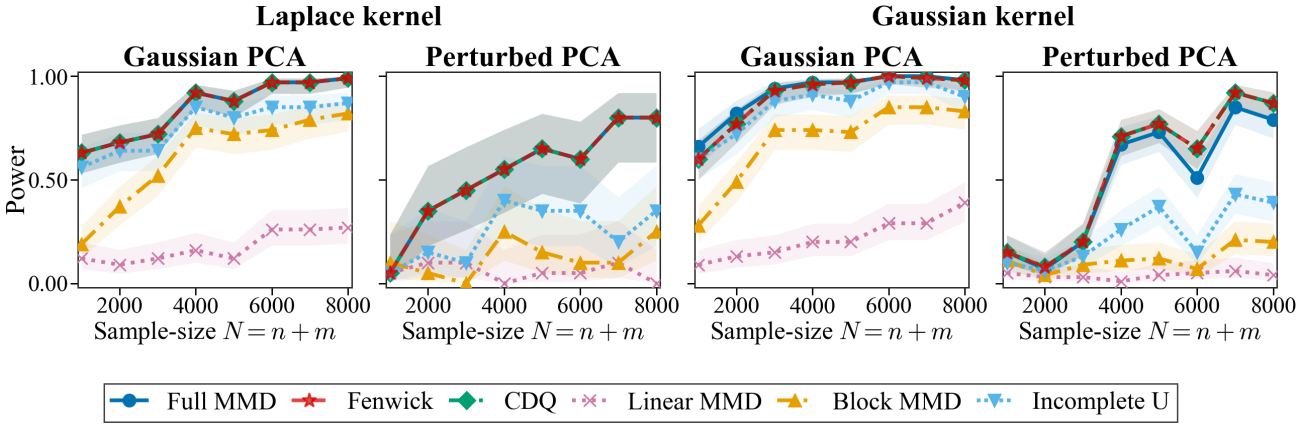

Figure 2. MMD permutation two sample testing experiments with dimension reduction via PCA.

the full pooled sample.

In Figure 2 shows PCA-MMD power versus sample size. In the Gaussian PCA setting, Fenwick and CDQ closely match Full MMD and reach high power quickly, while Linear MMD remains weak. In the perturbed PCA setting, Fenwick and CDQ again track Full MMD most closely, whereas Block MMD, Incomplete U-statistics, and Linear MMD lose substantial power.

Appendix F contains more experiments on MNIST data.

7. Conclusion

We propose a subquadratic MMD two-sample test that preserves the statistical strength of the full quadratic-time statistic. The method exploits orthantwise separability and Fenwick-tree prefix-sum queries to compute exact MMD sums for product Laplace-type kernels, and uses separable expansions to approximate Gaussian and inverse multiquadric kernels efficiently. With sufficiently accurate approximation, the resulting permutation test is finite-sample level-α and achieves the same minimax-rate optimality as full MMD in fixed or slowly growing dimension. Our algorithm relies on product or almost-separable kernels, and its complexity still grows exponentially in d through the num-

ber of orthants. It is also designed for offline computation, since rank compression requires access to the full dataset. For L_2 testing, the guarantees depend on the unknown s, although the aggregation method of Schrab et al. (2023) can be applied directly. Finally, the empirical evaluation is preliminary and should be expanded to more datasets, kernels, dimensions, and scalable MMD baselines.

Acknowledgements

The authors thank Arthur Gretton, Ilmun Kim, and Lester Mackey for beneficial discussions. PS and SP gratefully acknowledge NSF grants 2217069 and CCF-2505865. IC acknowledges partial support from the NIH Grant R01 NS121913.

Impact Statement

Our goal is to build and analyze subquadratic two-sample hypothesis tests with minimal distributional assumptions. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. Adaptive test of independence based on hsic measures. *The Annals of Statistics*, 50(2), 2022.

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

Bodenham, D. A. and Kawahara, Y. euMMD: efficiently computing the MMD two-sample test statistic for univariate data. *Statistics and Computing*, 33(5), 2023.

Butucea, C. Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5), 2007.

Chatalic, A., Schreuder, N., Rosasco, L., and Rudi, A. Nyström kernel mean embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3006–3024, 2022.

Chatalic, A., Letizia, M., Schreuder, N., and Rosasco, L. A scalable nystrom-based kernel two-sample test with permutations. *arXiv preprint arXiv:2502.13570*, 2025.

Choi, I. and Kim, I. Computational-statistical trade-off in kernel two-sample testing with random fourier features. *arXiv preprint arXiv:2407.08976*, 2024.

Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1981–1989, 2015.

Domingo-Enrich, C., Dwivedi, R., and Mackey, L. Compress then test: Powerful kernel testing in near-linear time. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1174–1218, 2023.

Dwivedi, R. and Mackey, L. Kernel thinning. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1753–1753, 2021.

Fenwick, P. M. A new data structure for cumulative frequency tables. *Software: Practice and Experience*, 24(3): 327–336, 1994.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1292–1301. JMLR.org, 2017.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. A fast, consistent kernel two-sample test. *Advances in neural information processing systems*, 22, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Hoeffding, W. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23(2):169–192, 1952.

Ingster, Y. I. Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

Kim, I. and Schrab, A. Differentially private permutation tests. *Journal of the American Statistical Association*, 2026.

Kim, I., Balakrishnan, S., and Wasserman, L. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1), 2022.

Kirchler, M., Khorasani, S., Kloft, M., and Lippert, C. Two-sample testing using deep learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1387–1398, 2020.

Langrené, N. and Warin, X. Fast multivariate empirical cumulative distribution function with connection to kernel density estimation. *Computational Statistics and Data Analysis*, 162:107267, 2021.

Li, T. and Yuan, M. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *Journal of Machine Learning Research*, 25(1), 2024.

Mishra, P. A new algorithm for updating and querying sub-arrays of multidimensional arrays. *arXiv preprint arXiv:1311.6093*, 2013.

Montgomery, H. L. and Vaughan, R. C. *Multiplicative Number Theory I: Classical Theory*. Cambridge University Press, 2006.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141, 2017.

Rudelson, M. and Vershynin, R. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. Efficient Aggregated Kernel Tests using Incomplete U-statistics. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18793–18807, 2022.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Simon-Gabriel, C.-J. and Schölkopf, B. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

Srivastava, N. and Vershynin, R. Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111, 2013.

Szabó, Z. and Sriperumbudur, B. K. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

Trefethen, L. N. and Weideman, J. The exponentially convergent trapezoidal rule. *SIAM review*, 56(3):385–458, 2014.

Wang, Y., Wu, Y., and Du, S. S. Near-linear time local polynomial nonparametric estimation with box kernels. *INFORMS Journal on Computing*, 2021.

Yamada, M., Wu, D., Tsai, Y. H., Ohta, H., Salakhutdinov, R., Takeuchi, I., and Fukumizu, K. Post selection inference with incomplete maximum mean discrepancy estimator. In *International Conference on Learning Representations*, 2019.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Zaremba, W., Gretton, A., and Blaschko, M. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 26, pp. 755–763, 2013.

Appendix roadmap.

1. Appendix A reviews related work on scalable kernel two-sample testing.
2. Appendix B gives the general minimax lower-bound framework used in the growing-dimensional analysis.
3. Appendix C contains the main proofs for theorems in Section 4 and Section 5, including kernel approximation, test validity, fixed-dimensional optimality, MMD-separation optimality, and growing-dimensional bounds.
4. Appendix D gives the full Fenwick-tree algorithms, including time and memory improvements via rank ordering, auxiliary ranks, and compressed trees.
5. Appendix E provides the PCA-MMD pseudocode and the proofs for the latent low-dimensional model.
6. Appendix F reports additional PCA-MMD experiments on MNIST.

A. Related work

The kernel two sample testing (Gretton et al., 2012) is known to be optimal from the minimax perspective (Kim & Schrab, 2026; Li & Yuan, 2024; Schrab et al., 2023), but it requires quadratic time complexity to compute its tests statistic due to its pairwise kernel evaluation. To address this computational challenge, several computationally efficient kernel based methods have been proposed (Gretton et al., 2012; Schrab et al., 2022; Yamada et al., 2019; Zaremba et al., 2013). However, these methods are either suboptimal in terms of testing power, or they eventually require quadratic time to achieve minimax optimality (Domingo-Enrich et al., 2023, Proposition 2). With this context in mind, whether kernel-based two sample testing can achieve minimax optimality within sub-quadratic time complexity has been remained unknown in the literature. We also note that there are more proposed computationally efficient methods (Chatalic et al., 2022, 2025; Chwialkowski et al., 2015; Kirchler et al., 2020) in the literature, yet their tradeoff between computational efficiency and testing power is not fully understood. Two notable exceptions are based on random Fourier features (Choi & Kim, 2024) and kernel thinning (Domingo-Enrich et al., 2023; Dwivedi & Mackey, 2021). They showed that sub-quadratic time is achievable for minimax optimality with respect to L_2-metric and MMD metric, respectively, under mild distributional assumptions such as bounded support, or sub-Gaussianity. In line with these works, we aim to expand sub-quadratic time guarantee for minimax optimality for more general distributions.

B. A framework for minimax lower bound analysis

In this section, we provide a common framework for deriving the lower bound for the minimax separation rate ϵ_n^*. We take a Bayesian approach, which is used as a standard tool in the literature to derive the lower bound for the minimax risk (Ingster, 1987; Butucea, 2007; Albert et al., 2022; Schrab et al., 2023; Li & Yuan, 2024). First, In particular, we recall Albert et al. (2022, Lemma 5), which is tailored for the two sample testing problem.

Lemma B.1 (Albert et al., 2022, Lemma 5)**.** *Let α, β in $(0,1)$ and η in $[0,1)$ be such that $\alpha+\beta+\eta<1$. Let us denote*

$$R^*(\epsilon_n) = \inf_{\varphi_n \in \Phi(\alpha)} \sup_{(P,Q)\in\mathcal{P}_1(\epsilon_n,\mathcal{C})} \mathbb{P}_{P\otimes Q}[\varphi_n = 0],$$

where $\Phi(\alpha) = \{\varphi_n : \sup_{(P,Q)\in\mathcal{P}_0} \mathbb{P}_{P\otimes Q}[\varphi_n = 1] \le \alpha\}$ is the set of all level-α tests. Let $\underline{\epsilon} > 0$ and consider a probability measure $\mu_{\underline{\epsilon}}$ defined on the set of paired distributions $\mathcal{P}(\mathbb{S})$ such that $\mu_{\underline{\epsilon}}(\mathcal{P}_1(\underline{\epsilon},\mathcal{C})) \ge 1-\eta$. Define the associated mixture measure for all measurable set A in $\mathbb{S}^n \times \mathbb{S}^m$ as

$$\mathbb{P}_{\mu_{\underline{\epsilon}}}(A) = \int_{\mathcal{P}(\mathbb{S})} \mathbb{P}_{P\otimes Q}(A) d\mu_{\underline{\epsilon}}(P,Q).$$

Assume there exists a distribution P_0 such that its pair with itself (P_0,P_0) is in $\mathcal{P}_0$, and the probability measure $\mu_{\underline{\epsilon}}$ is absolutely continuous with respect to $\mathbb{P}_{P_0\otimes P_0}$, and verifies

$$\mathbb{E}_{P_0\otimes P_0}\left[\left(\frac{d\mathbb{P}_{\mu_{\underline{\epsilon}}}(\mathbf{X},\mathbf{Y})}{d\mathbb{P}_{P_0\otimes P_0}(\mathbf{X},\mathbf{Y})}\right)^2\right] \le 1 + 4(1-\alpha-\beta-\eta)^2,$$

where $\frac{d\mathbb{P}_{\mu_{\underline{\epsilon}}}}{d\mathbb{P}_{P_0\otimes P_0}}$ *is the Radon-Nikodym derivative of* $\mathbb{P}_{\mu_{\underline{\epsilon}}}$ *with respect to* $\mathbb{P}_{P_0\otimes P_0}$*. Then, for all* $\epsilon \le \underline{\epsilon}$*, we have* $R^*(\epsilon_n) \ge \beta$*, which implies that*

$$\epsilon_n^* \ge \underline{\epsilon}.$$

We provide a proof sketch here. See Albert et al. (2022) for the original proof for independent testing case. By assumption $\mu_{\underline{\epsilon}}(\mathcal{P}_1(\underline{\epsilon}, \mathcal{C})) \ge 1-\eta$, we have

$$\begin{aligned}
\mathbb{P}_{\mu_{\underline{\epsilon}}}[\varphi_n = 0] &= \int_{\mathcal{P}(\mathbb{S})} \mathbb{P}_{P\otimes Q}[\varphi_n = 0] d\mu_{\underline{\epsilon}}(P,Q) \\
&\le \int_{\mathcal{P}_1(\underline{\epsilon},\mathcal{C})} \mathbb{P}_{P\otimes Q}[\varphi_n = 0] d\mu_{\underline{\epsilon}}(P,Q) + \mu_{\underline{\epsilon}}(\mathcal{P}_1^c(\underline{\epsilon},\mathcal{C})) \\
&\le \sup_{(P,Q)\in\mathcal{P}_1(\underline{\epsilon},\mathcal{C})} \mathbb{P}_{P\otimes Q}[\varphi_n = 0] + \eta,
\end{aligned}$$

which implies

$$\sup_{(P,Q)\in\mathcal{P}_1(\underline{\epsilon},\mathcal{C})} \mathbb{P}_{P\otimes Q}[\varphi_n = 0] \ge \mathbb{P}_{\mu_{\underline{\epsilon}}}[\varphi_n = 0] - \eta.$$

Therefore,

$$\begin{aligned}
R^*(\underline{\epsilon}) &= \inf_{\varphi_n\in\Phi(\alpha)} \sup_{(P,Q)\in\mathcal{P}_1(\underline{\epsilon},\mathcal{C})} \mathbb{P}_{P\otimes Q}[\varphi_n = 0] \\
&\ge \inf_{\varphi_n\in\Phi(\alpha)} \left\{ \mathbb{P}_{\mu_{\underline{\epsilon}}}[\varphi_n = 0] - \eta \right\} \\
&\ge 1 - \eta - \sup_{\varphi_n\in\Phi(\alpha)} \left\{ \mathbb{P}_{P_0\otimes P_0}[\varphi_n = 1] + \left(\mathbb{P}_{\mu_{\underline{\epsilon}}}[\varphi_n = 1] - \mathbb{P}_{P_0\otimes P_0}[\varphi_n = 1] \right) \right\} \\
&\ge 1 - \eta - \alpha - \sup_{\varphi_n\in\Phi(\alpha)} \left| \mathbb{P}_{P_0\otimes P_0}[\varphi_n = 1] - \mathbb{P}_{\mu_{\underline{\epsilon}}}[\varphi_n = 1] \right| \\
&\ge 1 - \eta - \alpha - \mathrm{TV}(\mathbb{P}_{\mu_{\underline{\epsilon}}}, \mathbb{P}_{P_0\otimes P_0}) \\
&\ge 1 - \eta - \alpha - \frac{1}{2} \left(\int \left(\frac{d\mathbb{P}_{\mu_{\underline{\epsilon}}}(\mathbf{x},\mathbf{y})}{d\mathbb{P}_{P_0\otimes P_0}(\mathbf{x},\mathbf{y})} \right)^2 d\mathbb{P}_{P_0\otimes P_0}(\mathbf{x},\mathbf{y}) - 1 \right)^{1/2} \\
&\ge \beta,
\end{aligned}$$

provided that

$$\mathbb{E}_{P_0\otimes P_0} \left[\left(\frac{d\mathbb{P}_{\mu_{\underline{\epsilon}}}(\mathbf{X},\mathbf{Y})}{d\mathbb{P}_{P_0\otimes P_0}(\mathbf{X},\mathbf{Y})} \right)^2 \right] \le 1 + 4(1-\alpha-\beta-\eta)^2. \tag{5}$$

Since $\mathcal{P}_1(\underline{\epsilon}, \mathcal{C}) \subseteq \mathcal{P}_1(\epsilon_n, \mathcal{C})$ for all $\epsilon_n \le \underline{\epsilon}$, we have $R^*(\epsilon_n) \ge R^*(\underline{\epsilon}) \ge \beta$ for all $\epsilon_n \le \underline{\epsilon}$, which implies $\epsilon_n^* \ge \underline{\epsilon}$.

Therefore, in order to derive the lower bound for ϵ_n^*, it suffices to find a distribution P_0 and a probability measure $\mu_{\underline{\epsilon}}$ satisfying the conditions in Lemma B.1. A straightforward way to construct such $\nu_{\underline{\epsilon}}$ is to consider a mixture over a family of paired distribution where one distribution is fixed as P_0 and the other distribution varies in a family of distributions that are separated from P_0 by at least $\underline{\epsilon}$ with respect to the distance of interest. In other words, we consider a family of paired distributions $\{(Q_{a_n}, P_0)\}_{a_n\in\Lambda_n}$, where Λ_n is an index set, and we construct $\mu_{\underline{\epsilon}}$ as assigning a prior distribution $\nu_{\underline{\epsilon}}$ on Λ_n. In that case, note that the second moment condition in Equation (6) can be simplified as follows. Suppose that Q_{a_n} is

absolutely continuous with respect to P_0 for each $a_n \in \Lambda_n$, and then

$$\begin{aligned}\frac{d\mathbb{P}_{Q_{a_n}\otimes P_0}}{d\mathbb{P}_{P_0\otimes P_0}}(\mathbf{X},\mathbf{Y}) &= \frac{d\left(Q_{a_n}^{\otimes n}\otimes P_0^{\otimes m}\right)}{d\left(P_0^{\otimes n}\otimes P_0^{\otimes m}\right)}(\mathbf{X},\mathbf{Y}) \\ &= \frac{dQ_{a_n}^{\otimes n}}{dP_0^{\otimes n}}(\mathbf{X})\frac{dP_0^{\otimes m}}{dP_0^{\otimes m}}(\mathbf{Y}) \\ &= \prod_{i=1}^{n}\frac{dQ_{a_n}}{dP_0}(X_i)\prod_{j=1}^{m}\frac{dP_0}{dP_0}(Y_j) \\ &= \prod_{i=1}^{n}\frac{dQ_{a_n}}{dP_0}(X_i).\end{aligned}$$

Therefore, if P_0 and Q_{a_n} have densities p_0 and q_{a_n} with respect to the Lebesgue measure, we obtain

$$\begin{aligned}\mathbb{E}_{P_0\otimes P_0}\left[\left(\frac{d\mathbb{P}_{\mu_{\underline{\epsilon}}}(\mathbf{X},\mathbf{Y})}{d\mathbb{P}_{P_0\otimes P_0}(\mathbf{X},\mathbf{Y})}\right)^2\right] &= \mathbb{E}_{P_0^{\otimes n}}\left[\left(\mathbb{E}_{A_n\sim\nu_{\underline{\epsilon}}}\left[\prod_{i=1}^{n}\frac{dQ_{A_n}}{dP_0}(X_i)\right]\right)^2\right] \\ &= \mathbb{E}_{\mathbf{p}_0}\left[\left(\frac{\mathbf{q}(\mathbf{X})}{\mathbf{p}_0(\mathbf{X})}\right)^2\right],\end{aligned}$$

where we denote $\mathbf{p}_0(\mathbf{x}) = \prod_{i=1}^{n} p_0(x_i)$ and $\mathbf{q}(\mathbf{x}) = \mathbb{E}_{A_n\sim\nu_{\underline{\epsilon}}}\left[\prod_{i=1}^{n} q_{A_n}(x_i)\right]$. Note that the above construction can be extended to the growing d setting in a straightforward manner, and the condition in Equation (5) is then simplified as

$$\mathbb{E}_{\mathbf{p}_0^{(n)}}\left[\left(\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})}\right)^2\right] \le 1 + 4(1-\alpha-\beta-\eta)^2. \tag{6}$$

General recipe for the construction of $\{q_{a_n}^{(n)}\}_{a_n\in\Lambda_n}$ Now, we introduce a general way of construction of $\{q_{a_n}^{(n)}\}_{a_n\in\Lambda_n}$. Given $p_0^{(n)}$ we construct $q_{a_n}^{(n)}$ as follows. First, assume there is a cube $U_d \in \mathbf{R}^d$ such that $\inf_{x\in U_d} p_0^{(n)}(x) \ge c_0 > 0$ for some constant c_0 that does not depend on n. For ease of analysis, we assume $U_d = [0,1]^d$. If such cube is $x_0 + l[0,1]^d$ for some $x_0 \in \mathbf{R}^d$ and $l > 0$, we can simply replace x by $(x - x_0)/l$ in the following construction.

For each n, choose an integer b_n, and let

$$\mathcal{K}_n = \{0,1,2,\ldots,b_n-1\}^d, \quad B_n := |\mathcal{K}_n| = b_n^d.$$

We partition U_d into B_n cubes

$$U_{n,k} = \prod_{j=1}^{d}\left(\frac{k_j}{b_n}, \frac{k_j+1}{b_n}\right), \quad k = (k_1,\ldots,k_d) \in \mathcal{K}_n.$$

Assumption B.2. Now, consider a family of measurable "bump" functions $\{\psi_{n,k}\}_{k\in\mathcal{K}_n}$ such that

1. $\mathrm{supp}(\psi_{n,k}) \subset U_{n,k}$ for each $k \in \mathcal{K}_n$.
2. $\int \psi_{n,k}(x)dx = 0$ for each $k \in \mathcal{K}_n$.
3. $\mathrm{supp}(\psi_{n,k})$ are disjoint for different $k \in \mathcal{K}_n$.
4. (Additional) the shape of $\psi_{n,k}$ for different k is the same, i.e., there exists a function ψ_n such that $\psi_{n,k}(x) = \psi_n(b_n x - k)$ for each $k \in \mathcal{K}_n$.

We will justify the existence of such family of functions later. Let $a_n = (a_{n,k})_{k\in\mathcal{K}_n} \in \{-1,1\}^{B_n}$, and we assign a prior ν_n on $\{-1,1\}^{B_n}$ as the uniform distribution, i.e., define $A_n = (A_{n,k})_{k\in\mathcal{K}_n}$, where $A_{n,k}$ are i.i.d. Rademacher random variables, and let ν_n be the distribution of A_n. We construct $q_{a_n}^{(n)}$ as

$$q_{a_n}^{(n)}(x) := p_0^{(n)}(x) + \omega_n \sum_{k\in\mathcal{K}_n} a_{n,k}\psi_{n,k}(x), \tag{7}$$

where $\omega_n > 0$ is a sequence of perturbation level that will be specified later. Then, define the n-sample product densities as

$$\mathbf{p}_0^{(n)}(\mathbf{x}^{(n)}) = \prod_{i=1}^{n} p_0^{(n)}(x_i^{(n)}), \quad \mathbf{q}_{a_n}^{(n)}(\mathbf{x}^{(n)}) = \prod_{i=1}^{n} q_{a_n}^{(n)}(x_i^{(n)}),$$

and the mixture distribution as

$$\mathbf{q}^{(n)}(\mathbf{x}^{(n)}) = \mathbb{E}_{A_n \sim \nu_n}\left[\mathbf{q}_{A_n}^{(n)}(\mathbf{x}^{(n)})\right] = 2^{-B_n} \sum_{a_n \in \{-1,1\}^{B_n}} \mathbf{q}_{A_n}^{(n)}(\mathbf{x}^{(n)}).$$

Further analysis of the χ^2-divergence bound Getting back to the χ^2-divergence bound in (6), we derive an upper bound for the left-hand side under the above construction.

Assumption B.3. Every $q_{a_n}^{(n)}$ defined in Equation (7) is a valid density function, i.e., $q_{a_n}^{(n)} \geq 0$ a.e. and $\int q_{a_n}^{(n)}(x)dx = 1$ for each $a_n \in \{-1,1\}^{B_n}$.

For each $k \in \mathcal{K}_n$, define

$$v_{n,k} := \int \frac{\psi_{n,k}(x)^2}{p_0^{(n)}(x)} dx.$$

Now, we claim that

$$\mathbb{E}_{\mathbf{p}_0^{(n)}}\left[\left(\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})}\right)^2\right] \leq \exp\left(\frac{1}{2} B_n n^2 \omega_n^4 \max_{k \in \mathcal{K}_n} v_{n,k}^2\right), \tag{8}$$

which have been shown in the previous literature, e.g., see Li & Yuan (2024); Ingster (1987). For completeness, we provide a proof of Equation (8) in the Appendix B.1.

$$1 + 4(1 - \alpha - \beta - \eta)^2$$

With Equation (6) in place, if we have

$$b_n^d n^2 \omega_n^4 \max_{k \in \mathcal{K}_n} v_{n,k}^2 \leq c(\alpha, \beta, \eta), \tag{9}$$

where $c(\alpha, \beta, \eta) = 2\log(1 + 4(1 - \alpha - \beta - \eta)^2)$, then we can conclude that $R^*(\epsilon_n) \geq \beta$. Therefore, the remain task is to choose $\{\psi_{n,k}\}_{k \in \mathcal{K}_n}$ which satisfy Assumption B.2 and Assumption B.3, so that $q_{a_n}^{(n)}$ are close to $p_0^{(n)}$ with respect to the distance of interest, and included in the prescribed class $\mathcal{P}_n$.

B.1. Proof of Equation (8)

Let us denote

$$\zeta_{n,k} := \frac{\psi_{n,k}(x)}{p_0^{(n)}(x)}.$$

Then,

$$\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})} = \mathbb{E}_{A_n \sim \nu_n}\left[\prod_{i=1}^{n}\left(1 + \omega_n \sum_{k \in \mathcal{K}_n} A_{n,k}\zeta_{n,k}(X_i^{(n)})\right)\right].$$

Note that

$$\mathbb{E}_{\mathbf{p}_0^{(n)}}\left[\left(\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})}\right)^2\right] = \mathbb{E}_{A_n, A_n' \sim \nu_n}\left[\mathbb{E}_{\mathbf{p}_0^{(n)}}\left\{\prod_{i=1}^{n}\left(1 + \omega_n \sum_{k \in \mathcal{K}_n} A_{n,k}\zeta_{n,k}(X_i^{(n)})\right)\left(1 + \omega_n \sum_{k \in \mathcal{K}_n} A_{n,k}'\zeta_{n,k}(X_i^{(n)})\right)\right\}\right],$$

and since $X_i^{(n)}$ are i.i.d. samples from $p_0^{(n)}$, we have

$$\mathbb{E}_{\mathbf{p}_0^{(n)}}\left[\left(\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})}\right)^2\right] = \mathbb{E}_{A_n, A_n' \sim \nu_n}\left[\left(\int\left(1 + \omega_n \sum_{k \in \mathcal{K}_n} A_{n,k}\zeta_{n,k}(x)\right)\left(1 + \omega_n \sum_{k \in \mathcal{K}_n} A_{n,k}'\zeta_{n,k}(x)\right)p_0^{(n)}(x)dx\right)^n\right].$$

Now, we focus on the inner expectation. Note that

$$\begin{aligned}
\int p_0^{(n)}(x)dx &= 1, \\
\int \zeta_{n,k}(x)p_0^{(n)}(x)dx &= \int \psi_{n,k}(x)dx = 0, \\
\int \zeta_{n,k}(x)\zeta_{n,k'}(x)p_0^{(n)}(x)dx &= \int \frac{\psi_{n,k}(x)\psi_{n,k'}(x)}{p_0^{(n)}(x)}dx = 0, \quad \text{for } k \neq k', \quad (\because \text{ supports are disjoint}) \\
\int \zeta_{n,k}(x)^2 p_0^{(n)}(x)dx &= \int \frac{\psi_{n,k}(x)^2}{p_0^{(n)}(x)}dx = v_{n,k}.
\end{aligned}$$

Therefore,

$$\int \left(1 + \omega_n \sum_{k\in\mathcal{K}_n} A_{n,k}\zeta_{n,k}(x)\right)\left(1 + \omega_n \sum_{k\in\mathcal{K}_n} A'_{n,k}\zeta_{n,k}(x)\right)p_0^{(n)}(x)dx = 1 + \omega_n^2 \sum_{k\in\mathcal{K}_n} A_{n,k}A'_{n,k}v_{n,k},$$

and this implies

$$\mathbb{E}_{\mathbf{p}_0^{(n)}}\left[\left(\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})}\right)^2\right] = \mathbb{E}_{A_n, A'_n\sim\nu_n}\left[\left(1 + \omega_n^2 \sum_{k\in\mathcal{K}_n} A_{n,k}A'_{n,k}v_{n,k}\right)^n\right].$$

Observe that

$$1 + \omega_n^2 \sum_{k\in\mathcal{K}_n} A_{n,k}A'_{n,k}v_{n,k} = \int \frac{q_{A_n}^{(n)}(x)q_{A'_n}^{(n)}(x)}{p_0^{(n)}(x)^2}p_0^{(n)}(x)dx \geq 0,$$

since we assume $q_{a_n}^{(n)}$ is a valid density function for each $a_n \in \{-1,1\}^{B_n}$. Therefore, we can apply the inequality $1+x \leq e^x$ for $x \geq 0$, and get

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}_0^{(n)}}\left[\left(\frac{\mathbf{q}^{(n)}(\mathbf{X}^{(n)})}{\mathbf{p}_0^{(n)}(\mathbf{X}^{(n)})}\right)^2\right] &\leq \mathbb{E}_{A_n, A'_n\sim\nu_n}\left[\exp\left(n\omega_n^2 \sum_{k\in\mathcal{K}_n} A_{n,k}A'_{n,k}v_{n,k}\right)\right] \\
&= \mathbb{E}_{\tilde{A}_{n,k}}\left[\exp\left(n\omega_n^2 \sum_{k\in\mathcal{K}_n} \tilde{A}_{n,k}v_{n,k}\right)\right],
\end{aligned}$$

where $\tilde{A}_{n,k} = A_{n,k}A'_{n,k}$ for each $k \in \mathcal{K}_n$. Observe that $\tilde{A}_{n,k} = A_{n,k}A'_{n,k}$ are again i.i.d. Rademacher random variables, and thus

$$\begin{aligned}
\mathbb{E}_{\tilde{A}_{n,k}}\left[\exp\left(n\omega_n^2 \sum_{k\in\mathcal{K}_n} \tilde{A}_{n,k}v_{n,k}\right)\right] &= \prod_{k\in\mathcal{K}_n} \mathbb{E}_{\tilde{A}_{n,k}}\left[\exp\left(n\omega_n^2 \tilde{A}_{n,k}v_{n,k}\right)\right] \\
&= \prod_{k\in\mathcal{K}_n} \cosh(n\omega_n^2 v_{n,k}) \\
&\leq \prod_{k\in\mathcal{K}_n} \exp\left(\frac{1}{2}n^2\omega_n^4 v_{n,k}^2\right) \\
&\leq \exp\left(\frac{1}{2}B_n n^2 \omega_n^4 \max_{k\in\mathcal{K}_n} v_{n,k}^2\right),
\end{aligned}$$

where the first equality holds since $\tilde{A}_{n,k}$ are independent, the second equality holds since $\tilde{A}_{n,k}$ are Rademacher, and the inequality holds since $\cosh(x) \leq e^{x^2/2}$ for each $x \in \mathbb{R}$.

C. Proofs of Theorems in Section 4 and Section 5

C.1. Proof of Theorem 3.3

Now, we show the approximation error bound. Let us denote $\tilde{k}_i = \sum_{r_i=1}^{R_i} \tilde{k}_{i,r_i}$ as the (R_i, δ_i)-APS approximation of k_i for each $i = 1, \dots, d$. We have Note that we have the telescopic decompsition

$$\prod_{i=1}^{d} a_i - \prod_{i=1}^{d} b_i = \sum_{i=1}^{d} (a_i - b_i) \Big(\prod_{j=1}^{i-1} a_j \Big) \Big(\prod_{j=i+1}^{d} b_j \Big),$$

and thus

$$\begin{aligned} \Big| \prod_{i=1}^{d} a_i - \prod_{i=1}^{d} b_i \Big| &\leq \sum_{i=1}^{d} |a_i - b_i| \Big(\prod_{j=1}^{i-1} |a_j| \Big) \Big(\prod_{j=i+1}^{d} |b_j| \Big) \\ &\leq \sum_{i=1}^{d} |a_i - b_i| \Big(\prod_{j=1}^{i-1} |a_j| \Big) \Big(\prod_{j=i+1}^{d} (|a_j| + |a_j - b_j|) \Big) \end{aligned}$$

holds for any $a_i, b_i \in \mathbb{R}$ for $i = 1, \dots, d$. Plugging in $a_i = k_i(x_i, y_i)$ and $b_i = \tilde{k}_i(x_i, y_i)$, we have

$$\begin{aligned} \Big| \prod_{i=1}^{d} k_i(x_i, y_i) - \prod_{i=1}^{d} \tilde{k}_i(x_i, y_i) \Big| &\leq \sum_{i=1}^{d} \delta_i \Big(\prod_{j=1}^{i-1} U_j \Big) \Big(\prod_{j=i+1}^{d} (U_j + \delta_j) \Big) \\ &= \sum_{i \in A} \delta_i \Big(\prod_{j=1}^{i-1} U_j \Big) \Big(\prod_{j=i+1}^{d} (U_j + \delta_j) \Big) \\ &\leq \sum_{i \in A} \delta_i \Big(\prod_{j \neq i} (U_j + \delta_j) \Big) \\ &= \sum_{i \in A} \delta_i \Big(\prod_{j \neq i, j \in A} (U_{\max} + \delta_{\max}) \prod_{j \neq i, j \in A^c} U_j \Big) \\ &= \sum_{i \in A} \delta_i (U_{\max} + \delta_{\max})^{m-1} \prod_{j \in A^c} U_j \\ &\leq \Big(\prod_{j \in A^c} U_j \Big) m \delta_{\max} (U_{\max} + \delta_{\max})^{m-1}. \end{aligned}$$

Lastly, we check $\tilde{k}$ is indeed a sum of $\prod_{i=1}^{d} R_i$ many OS kernels. Note that

$$\prod_{i=1}^{d} \tilde{k}_i(x_i, y_i) = \prod_{i=1}^{d} \Big(\sum_{r_i=1}^{R_i} \tilde{k}_{i,r_i}(x_i, y_i) \Big) = \sum_{r_1=1}^{R_1} \cdots \sum_{r_d=1}^{R_d} \prod_{i=1}^{d} \tilde{k}_{i,r_i}(x_i, y_i),$$

and thus it suffices to check that $\prod_{i=1}^{d} \tilde{k}_{i,r_i}(x_i, y_i)$ is an OS kernel for any $r_1, \dots, r_d$. Observe that for each $r_1, \dots, r_d$ and each i, $\tilde{k}_{i,r_i}(x_i, y_i)$ is PS kernel, and thus there exists a function $u_{+,r_i}, u_{-,r_i}, v_{+,r_i}, v_{-,r_i}$ defined on $\mathbb{R}$ such that

$$\tilde{k}_{i,r_i}(x_i, y_i) = \mathbf{1}\{y_i \leq x_i\} u_{+,r_i}(x_i) v_{+,r_i}(y_i) + \mathbf{1}\{y_i > x_i\} u_{-,r_i}(x_i) v_{-,r_i}(y_i).$$

Consider $s = (s_1, \ldots, s_d) \in \{+,-\}^d$, then we have

$$\begin{aligned}
&\prod_{i=1}^d \tilde{k}_{i,r_i}(x_i, y_i) \\
&= \prod_{i=1}^d \Big(\mathbf{1}\{y_i \le x_i\} u_{+,r_i}(x_i) v_{+,r_i}(y_i) + \mathbf{1}\{y_i > x_i\} u_{-,r_i}(x_i) v_{-,r_i}(y_i) \Big) \\
&= \prod_{i=1}^d \Big(\mathbf{1}\{\mathrm{sign}(y_i - x_i) = -\} u_{+,r_i}(x_i) v_{+,r_i}(y_i) + \mathbf{1}\{\mathrm{sign}(y_i - x_i) = +\}(y_i - x_i) u_{-,r_i}(x_i) v_{-,r_i}(y_i) \Big) \\
&= \sum_{s \in \{+,-\}^d} \prod_{i=1}^d \Big(\mathbf{1}\{\mathrm{sign}(y_i - x_i) = s_i\} u_{s_i^-,r_i}(x_i) v_{s_i^-,r_i}(y_i) \Big) \\
&= \sum_{s \in \{+,-\}^d} \mathbf{1}\{\mathrm{sign}(y_i - x_i) = s_i, \forall i\} \Big(\prod_{i=1}^d u_{s_i^-,r_i}(x_i) \Big) \prod_{i=1}^d \Big(v_{s_i^-,r_i}(y_i) \Big),
\end{aligned}$$

where we use the notation s_i^- to denote the opposite sign of s_i. We complete the proof by taking

$$f_s(x) = \prod_{i=1}^d u_{s_i^-,r_i}(x_i), \quad g_s(y) = \prod_{i=1}^d v_{s_i^-,r_i}(y_i).$$

C.2. Details on Table 1

In this section we provide a detailed lemma which expands on Table 1 with general δ.

Lemma C.1 (APS kernel approximation of characteristic kernels)**.** *The following characteristic kernels allows APS approximation with the following (R, δ):*

1. *Laplace kernel: $k(x,y) = \frac{1}{2} e^{-|x-y|}$ is $(1,0)$-APS kernel.*
2. *Matérn kernel with half-integer parameter $\nu = p + 1/2$: $k(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \big(\sqrt{2\nu}|x-y|\big)^\nu K_\nu\big(\sqrt{2\nu}|x-y|\big)$ is $(p+1,0)$-APS kernel, where p is a non-negative integer, and K_ν is the modified Bessel function of the second kind.*
3. *Gaussian kernel: $k(x,y) = \frac{1}{\sqrt{\pi}} e^{-|x-y|^2}$ is $\big(\lceil C_{\delta_0} \log(1/\delta) \rceil,\ \delta\big)$-APS kernel for any $\delta \in (0, \delta_0]$, where $\delta_0 \in (0,1)$ is a fixed constant and $C_{\delta_0} > 0$ is a constant that depends on δ_0.*
4. *Inverse multiquadric kernel: $k(x,y) = (1 + |x-y|^2/c^2)^{-\gamma}$ is $\big(\lceil C_{\gamma,\delta_0} \log^3(1/\delta) \rceil,\ \delta\big)$-APS kernel for any $\delta \in (0, \delta_0]$, where $\delta_0 \in (0,1)$ is a fixed constant and $C_{\gamma,\delta_0} > 0$ is a constant that depends on $\gamma > 0$ and $\delta_0 \in (0,1)$.*

C.3. Proof of Lemma C.1

C.3.1. Matérn kernel with half-integer parameter $\nu = p + 1/2$

Proof. Write $r = |x-y|$ and $\lambda = \sqrt{2\nu} = \sqrt{2p+1}$. For half-integer smoothness $\nu = p + 1/2$, the Matérn kernel admits the closed-form representation

$$k_\nu(x,y) = e^{-\lambda r} P_p(\lambda r),$$

where $P_p(t) = \frac{p!}{(2p)!} \sum_{j=0}^p \frac{(p+j)!}{j!(p-j)!} (2t)^{p-j}$ is a polynomial of degree p. Hence there exist constants $c_0, \ldots, c_p$ such that

$$k_\nu(x,y) = e^{-\lambda|x-y|} \sum_{q=0}^p c_q |x-y|^q.$$

We show that this kernel can be written as a sum of $p+1$ piecewise separable kernels. On the region $y \leq x$, we have $|x-y| = x-y$, and therefore

$$\begin{aligned} k_\nu(x,y) &= e^{-\lambda(x-y)} \sum_{q=0}^{p} c_q (x-y)^q \\ &= e^{-\lambda x} e^{\lambda y} \sum_{q=0}^{p} c_q \sum_{\ell=0}^{q} \binom{q}{\ell} x^{q-\ell} (-y)^\ell \\ &= \sum_{\ell=0}^{p} \left[e^{-\lambda x} \sum_{q=\ell}^{p} c_q \binom{q}{\ell} (-1)^\ell x^{q-\ell} \right] e^{\lambda y} y^\ell. \end{aligned}$$

For each $\ell = 0, \ldots, p$, define

$$f_{\ell,-}(x) = e^{-\lambda x} \sum_{q=\ell}^{p} c_q \binom{q}{\ell} (-1)^\ell x^{q-\ell}, \qquad g_{\ell,-}(y) = e^{\lambda y} y^\ell,$$

Similarly, on the region $\{y > x\}$, we have $|x-y| = y-x$, and hence

$$\begin{aligned} k_\nu(x,y) &= e^{-\lambda(y-x)} \sum_{q=0}^{p} c_q (y-x)^q \\ &= e^{\lambda x} e^{-\lambda y} \sum_{q=0}^{p} c_q \sum_{\ell=0}^{q} \binom{q}{\ell} y^{q-\ell} (-x)^\ell \\ &= \sum_{\ell=0}^{p} \left[e^{\lambda x} (-x)^\ell \right] \left[e^{-\lambda y} \sum_{q=\ell}^{p} c_q \binom{q}{\ell} y^{q-\ell} \right]. \end{aligned}$$

For each $\ell = 0, \ldots, p$, define

$$f_{\ell,+}(x) = e^{\lambda x} (-x)^\ell, \qquad g_{\ell,+}(y) = e^{-\lambda y} \sum_{q=\ell}^{p} c_q \binom{q}{\ell} y^{q-\ell}.$$

Define

$$k_l(x,y) = \mathbf{1}\{y \leq x\} f_{\ell,-}(x) g_{\ell,-}(y) + \mathbf{1}\{y > x\} f_{\ell,+}(x) g_{\ell,+}(y),$$

for each $\ell = 0, \ldots, p$. We observe that $k_\nu(x,y) = \sum_{\ell=0}^{p} k_\ell(x,y)$ and completes the proof.

□

C.3.2. GAUSSIAN KERNEL

We use the trapezoidal rule (Trefethen & Weideman, 2014) to approximate the Gaussian kernel. First, observe that the following lemma holds.

Lemma C.2. *Let $\rho \geq 0$. For $a > 0$, define*

$$g_\rho(u) = \frac{1}{\sqrt{\pi}} \exp\big((a+iu)^2 - 2(a+iu)\rho\big). \tag{10}$$

Then

$$\int_{\mathbb{R}} g_\rho(u)\, du = e^{-\rho^2}.$$

Moreover, for $T \geq 1$, consider the approximation

$$I_T(\rho) = \frac{1}{T} \sum_{j \in \mathbb{Z}} g_\rho\Big(\frac{j}{T}\Big), \tag{11}$$

then

$$\sup_{\rho \geq 0} \big| e^{-\rho^2} - I_T(\rho) \big| \leq 2e^{-\pi^2 T^2/2}.$$

Proof. First, note that

$$g_\rho(u) = \frac{1}{\sqrt{\pi}} e^{a^2 - 2a\rho} e^{-u^2} e^{2iu(a-\rho)}. \tag{12}$$

Hence

$$\begin{aligned}
\int_{\mathbb{R}} g_\rho(u)\, du &= e^{a^2-2a\rho} \int_{\mathbb{R}} e^{2iu(a-\rho)} \Big(\frac{1}{\sqrt{\pi}} e^{-u^2} \Big) du \\
&= e^{a^2-2a\rho} \mathbb{E}_{U \sim N(0,1/2)} \Big[e^{2iU(a-\rho)} \Big] \\
&= e^{a^2-2a\rho} e^{-(a-\rho)^2} \quad \big(\because \mathbb{E}_{X \sim N(0,\sigma^2)} [e^{itX}] = e^{-\sigma^2 t^2/2} \big) \\
&= e^{-\rho^2}.
\end{aligned}$$

Similarly, note that

$$\begin{aligned}
\widehat{g}_\rho(\xi) &= \frac{1}{\sqrt{\pi}} e^{a^2-2a\rho} \int_{\mathbb{R}} e^{-u^2} e^{iu(2a-2\rho-\xi)} du \\
&= e^{a^2-2a\rho} \int_{\mathbb{R}} e^{iu(2a-2\rho-\xi)} \Big(\frac{1}{\sqrt{\pi}} e^{-u^2} \Big) du \\
&= e^{a^2-2a\rho} e^{-(2a-2\rho-\xi)^2/4} \\
&= \exp\Big(-\rho^2 + \xi(a-\rho) - \frac{\xi^2}{4} \Big)
\end{aligned}$$

By Fourier inversion,

$$g_\rho(u) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{g}_\rho(\xi) e^{iu\xi}\, d\xi.$$

Thus

$$\begin{aligned}
I_T(\rho) &= \frac{1}{T} \sum_{j \in \mathbb{Z}} g_\rho\Big(\frac{j}{T}\Big) \\
&= \frac{1}{2\pi T} \int_{\mathbb{R}} \widehat{g}_\rho(\xi) \sum_{j \in \mathbb{Z}} e^{i\xi j/T}\, d\xi \\
&= \sum_{\ell \in \mathbb{Z}} \widehat{g}_\rho(2\pi T \ell),
\end{aligned}$$

where the last step uses the Poisson sum distribution identity:

$$\sum_{j \in \mathbb{Z}} e^{i\xi j/T} = 2\pi T \sum_{\ell \in \mathbb{Z}} \delta(\xi - 2\pi T \ell)$$

where δ is the Dirac delta distribution. Since $\widehat{g}_\rho(0) = e^{-\rho^2}$, the discretization error is

$$\begin{aligned}
&I_T(\rho) - e^{-\rho^2} \\
&= \sum_{\ell \neq 0} \widehat{g}_\rho(2\pi T\ell) \\
&= \sum_{\ell \geq 1} \widehat{g}_\rho(2\pi T\ell) + \sum_{\ell \leq -1} \widehat{g}_\rho(2\pi T\ell) \\
&= \sum_{\ell \geq 1} \widehat{g}_\rho(2\pi T\ell) + \sum_{p \geq 1} \widehat{g}_\rho(-2\pi Tp) \\
&= \sum_{\ell \geq 1} \exp\Big(-\rho^2 + 2\pi T\ell(a-\rho) - \pi^2 T^2 \ell^2 \Big) + \sum_{p \geq 1} \exp\Big(-\rho^2 - 2\pi Tp(a-\rho) - \pi^2 T^2 p^2 \Big) \\
&\leq \sum_{\ell \geq 1} \exp\Big(2\pi T\ell a - \pi^2 T^2 \ell^2 \Big) + \sum_{p \geq 1} \exp\Big(-2\pi Tpa \Big), \quad (\because -(\rho + \pi Tp)^2 \leq 0)
\end{aligned}$$

Now set $a = \pi T/4$. Then, since $l(l-1/2) \geq l/2$ for all $l \geq 1$,

$$\begin{aligned}
\sum_{\ell \geq 1} \exp\Big(2\pi T\ell a - \pi^2 T^2 \ell^2 \Big) &= \sum_{\ell \geq 1} \exp\Big(-\pi^2 T^2 \ell(\ell - 1/2) \Big) \\
&\leq \sum_{\ell \geq 1} \exp\Big(-\frac{\pi^2 T^2 \ell}{2} \Big) \\
&= \frac{e^{-\pi^2 T^2/2}}{1 - e^{-\pi^2 T^2/2}}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sum_{p \geq 1} \exp\Big(-2\pi Tpa \Big) &= \sum_{p \geq 1} \exp\Big(-\pi^2 T^2 p/2 \Big) \\
&= \frac{e^{-\pi^2 T^2/2}}{1 - e^{-\pi^2 T^2/2}}
\end{aligned}$$

Therefore, we combine the two bounds to get

$$\sup_{\rho \geq 0} \left| e^{-\rho^2} - I_T(\rho) \right| \leq \frac{2q}{1-q}, \quad \text{where } q = e^{-\pi^2 T^2/2},$$

which is bounded by $3q$, since we have $q \leq 1/3$ from the assumption $T \geq 1$.

□

Lemma C.3. *Let $\rho \geq 0$, $T \geq 1$, and $M > 0$. Assume for simplicity that $MT \in \mathbb{N}$, and define*

$$I_{T,M}(\rho) = \frac{1}{T} \sum_{|j| \leq MT} g_\rho\Big(\frac{j}{T} \Big).$$

If $M \geq \frac{3}{4}\pi T$, then

$$\sup_{\rho \geq 0} \left| e^{-\rho^2} - I_{T,M}(\rho) \right| \leq 3e^{-\pi^2 T^2/2}.$$

Proof. We decompose

$$\left| e^{-\rho^2} - I_{T,M}(\rho) \right| \leq \left| e^{-\rho^2} - I_T(\rho) \right| + \left| I_T(\rho) - I_{T,M}(\rho) \right|.$$

By Lemma C.2,

$$\sup_{\rho \geq 0} \left| e^{-\rho^2} - I_T(\rho) \right| \leq 2e^{-\pi^2 T^2/2}.$$

It remains to bound the truncation error arised from M. Recall Equation (12):

$$g_\rho(u) = \frac{1}{\sqrt{\pi}} e^{a^2 - 2a\rho} e^{-u^2} e^{2iu(a-\rho)} = e^{-2a\rho} \cdot \frac{e^{a^2}}{\sqrt{\pi}} e^{-u^2} e^{2iu(a-\rho)}.$$

We have

$$\sup_{\rho \geq 0} |g_\rho(u)| \leq \frac{e^{a^2}}{\sqrt{\pi}} e^{-u^2},$$

provided that $a\rho \geq 0$. Therefore

$$\left| I_T(\rho) - I_{T,M}(\rho) \right| \leq \frac{e^{a^2}}{T\sqrt{\pi}} \sum_{|j| > MT} e^{-(j/T)^2}.$$

Since $u \mapsto e^{-u^2}$ is decreasing on $[0, \infty)$,

$$\frac{1}{T} \sum_{|j| > MT} e^{-(j/T)^2} \leq 2 \int_M^\infty e^{-u^2} \, du.$$

Observe that we have the inequality

$$\int_M^\infty e^{-u^2} \, du \leq \frac{1}{M} \int_M^\infty u e^{-u^2} \, du \leq \frac{e^{-M^2}}{2M}.$$

Therefore, the above analysis gives

$$\begin{aligned} \left| I_T(\rho) - I_{T,M}(\rho) \right| &\leq \frac{e^{a^2}}{\sqrt{\pi}} \cdot \frac{1}{T} \sum_{|j| > MT} e^{-(j/T)^2} \\ &\leq \frac{e^{a^2 - M^2}}{\sqrt{\pi} M}. \end{aligned}$$

Note that the above function is decreasing with respect to M. With $a = \pi T/4$ and $M \geq \frac{3\pi}{4} T$,

$$\begin{aligned} \left| I_T(\rho) - I_{T,M}(\rho) \right| &\leq \frac{4}{3\pi^{3/2} T} e^{-\pi^2 T^2/2} \\ &\leq e^{-\pi^2 T^2/2}, \end{aligned}$$

Combining the two estimates gives

$$\sup_{\rho \geq 0} \left| e^{-\rho^2} - I_{T,M}(\rho) \right| \leq 3e^{-\pi^2 T^2/2}.$$

□

Now, we show that $I_{T,M}(|x-y|)$ is a sum of $2L+1$ real piecewise separable kernels. Recall Equation (10):

$$g_\rho(u) = \frac{1}{\sqrt{\pi}} \exp\big((a+iu)^2 - 2(a+iu)\rho\big) \tag{13}$$

$$= \frac{1}{\sqrt{\pi}} e^{a^2 - 2a\rho} e^{-u^2} e^{2iu(a-\rho)}, \tag{14}$$

and the definition of the truncated approximation:

$$I_{T,M}(\rho) = \frac{1}{T} \sum_{|j| \le MT} g_\rho\Big(\frac{j}{T}\Big).$$

Note that our choice is $a = \pi T/4$, and $M \ge \frac{3}{4}\pi T$. Let us define $L = MT \in \mathbb{N}$, and $u_j = j/T$. Pairing the j and $-j$ terms in the truncated trapezoidal approximation gives the following equivalent real-valued expression:

$$\begin{aligned} I_{T,M}(\rho) &= \frac{e^{a^2-2a\rho}}{T\sqrt{\pi}}\Big(1 + 2\sum_{j=1}^{L} e^{-u_j^2}\cos\big(2u_j(a-\rho)\big)\Big) \\ &= \frac{e^{a^2}}{T\sqrt{\pi}}e^{-2a\rho} + \sum_{j=1}^{L}\frac{2e^{a^2-u_j^2}}{T\sqrt{\pi}}e^{-2a\rho}\cos\big(2u_j(a-\rho)\big). \end{aligned}$$

Define

$$A_0 = \frac{e^{a^2}}{T\sqrt{\pi}}, \qquad A_j = \frac{2e^{a^2-u_j^2}}{T\sqrt{\pi}}, \qquad j = 1, \dots, L.$$

With $\rho = |x-y|$, we rewrite $I_{T,M}(\rho)$ as

$$I_{T,M}(\rho) = h_0(x,y) + \sum_{j=1}^{L} h_j(x,y),$$

where

$$h_0(x,y) = A_0 e^{-2a|x-y|}, \quad h_j(x,y) = A_j e^{-2a|x-y|}\cos\big(2u_j(a-|x-y|)\big),$$

for $j = 1, \dots, L$.

The first term can be easily seen to be a real piecewise separable kernel:

$$h_0(x,y) = \begin{cases} A_0 e^{-2ax}e^{2ay}, & x \ge y, \\ A_0 e^{2ax}e^{-2ay}, & x < y. \end{cases}$$

For the second term, we have two cases:

$$h_j(x,y) = \begin{cases} A_j e^{-2ax}e^{2ay}\cos\big(2u_j(a-x+y)\big), & x \ge y, \\ A_j e^{2ax}e^{-2ay}\cos\big(2u_j(a+x-y)\big), & x < y. \end{cases}$$

On $x \ge y$, using $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, this can be written as

$$h_j(x,y) = A_j e^{-2ax}\cos\big(2u_j(a-x)\big)e^{2ay}\cos(2u_j y) - A_j e^{-2ax}\sin\big(2u_j(a-x)\big)e^{2ay}\sin(2u_j y).$$

On $x < y$, using $\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$, this becomes

$$h_j(x,y) = A_j e^{2ax}\cos\big(2u_j(a+x)\big)e^{-2ay}\cos(2u_j y) + A_j e^{2ax}\sin\big(2u_j(a+x)\big)e^{-2ay}\sin(2u_j y).$$

Therefore, for each $j = 1, \dots, L$, h_j is the sum of two real piecewise separable kernels. More explicitly, write

$$h_j(x,y) = h_{j,c}(x,y) + h_{j,s}(x,y),$$

where

$$h_{j,c}(x,y) = \begin{cases} A_j e^{-2ax}\cos\big(2u_j(a-x)\big)e^{2ay}\cos(2u_j y), & x \ge y, \\ A_j e^{2ax}\cos\big(2u_j(a+x)\big)e^{-2ay}\cos(2u_j y), & x < y, \end{cases}$$

and

$$h_{j,s}(x,y) = \begin{cases} -A_j e^{-2ax}\sin\big(2u_j(a-x)\big)e^{2ay}\sin(2u_j y), & x \geq y, \\ A_j e^{2ax}\sin\big(2u_j(a+x)\big)e^{-2ay}\sin(2u_j y), & x < y. \end{cases}$$

Thus $I_{T,M}(\rho)$ is the sum of $1 + 2L = 2L + 1$ real piecewise separable kernels.

Now, we summarize the above analysis. First, by Lemma C.3, we have

$$\sup_{\rho \geq 0} \left| e^{-\rho^2} - I_T(\rho) \right| \leq 3e^{-\pi^2 T^2/2}.$$

with the choice of $M \geq \frac{3}{4}\pi T$, by Lemma C.3, and we showed that $I_{T,M}(\rho)$ is a sum of $2L+1$ real piecewise separable kernels, where $L = MT$. The gaussian kernel is $k(x,y) = \frac{1}{\sqrt{\pi}} e^{-(x-y)^2}$, and we take $\tilde{k}(x,y) = \frac{1}{\sqrt{\pi}} I_{T,M}(|x-y|)$ as the approximation. Note that the above analysis gives

$$\sup_{x,y \in \mathbb{R}} \left| k(x,y) - \frac{1}{\sqrt{\pi}} I_{T,M}(|x-y|) \right| \leq \frac{3}{\sqrt{\pi}} e^{-\pi^2 T^2/2}.$$

Given a target error $\delta > 0$, if suffices to choose T such that

$$\frac{3}{\sqrt{\pi}} e^{-\pi^2 T^2/2} \leq \delta.$$

A sufficient condition is

$$T^2 \geq \max\left\{1, \frac{2}{\pi^2}\log\left(\frac{3}{\sqrt{\pi}\delta}\right)\right\}.$$

where the first term is from the assumption $T \geq 1$. With this choice of T, we have Let $M = \frac{3}{4}\pi T$, then it is equivalent to choose $L \geq \max\{\frac{3\pi}{4}, \frac{3}{2\pi}\log(3/\delta)\}$. Considering L should be an integer, we choose

$$L = \left\lceil \max\left\{\frac{3\pi}{4}, \frac{3}{2\pi}\log\left(\frac{3}{\sqrt{\pi}\delta}\right)\right\}\right\rceil.$$

This implies that in order to achieve an approximation error of at most δ, it suffices to use

$$R = 2\left\lceil \max\left\{\frac{3\pi}{4}, \frac{3}{2\pi}\log\left(\frac{3}{\sqrt{\pi}\delta}\right)\right\}\right\rceil + 1$$

real piecewise separable kernels in the approximation. Furthermore, if $0 < \delta \leq \delta_0 < 1$, note that

$$\begin{aligned} R &\leq \frac{3\pi}{2} + \frac{3}{\pi}\log\left(\frac{3}{\sqrt{\pi}\delta}\right) + 2 \\ &\leq \frac{3\pi}{2} + \frac{3}{\pi}\log\left(\frac{3}{\sqrt{\pi}}\right) + 2 + \frac{3}{\pi}\log\left(\frac{1}{\delta}\right) \\ &\leq \left[\frac{3}{\pi} + \frac{\frac{3\pi}{2} + \frac{3}{\pi}\log(3/\sqrt{\pi}) + 2}{\log(1/\delta)}\right]\log\left(\frac{1}{\delta}\right) \\ &\leq \left[\frac{3}{\pi} + \frac{\frac{3\pi}{2} + \frac{3}{\pi}\log(3/\sqrt{\pi}) + 2}{\log(1/\delta_0)}\right]\log\left(\frac{1}{\delta}\right) \\ &\leq c(\delta_0)\log\left(\frac{1}{\delta}\right). \end{aligned}$$

For example, if $\delta_0 = 0.1$, then we note that $c(\delta_0) \leq 5$.

C.3.3. INVERSE MULTIQUADRIC KERNEL

A proof procedure is similar to Appendix C.3.2. Let

$$K_{\gamma,c}(u) = \left(1 + \frac{u^2}{c^2}\right)^{-\gamma},$$

for $c > 0$ and $\gamma > 0$. Let us denote $A := 1 + \frac{u^2}{c^2}$. By Gamma function identity, we have

$$A^{-\gamma} = \frac{1}{\Gamma(\gamma)} \int_0^\infty t^{\gamma-1} e^{-tA}\, dt,$$

which implies that

$$\begin{aligned} K_{\gamma,c}(u) &= \frac{1}{\Gamma(\gamma)} \int_0^\infty t^{\gamma-1} e^{-t} e^{-\frac{t}{c^2}u^2} dt \\ &= \frac{1}{\Gamma(\gamma)} \int_{-\infty}^\infty e^{\gamma s - e^s} e^{-u^2 e^s/c^2} ds \quad \text{(by taking } t = e^s\text{)}. \end{aligned}$$

Now, we apply the trapezoidal approximation to the above integral. For $T \geq 1$, define

$$I_T(u) = \frac{1}{\Gamma(\gamma)T} \sum_{j\in\mathbb{Z}} e^{\gamma j/T - e^{j/T}} e^{-e^{j/T}u^2/c^2},$$

and observe that this is gaussian mixture,

$$I_T(u) = \sum_{j\in\mathbb{Z}} w_j e^{-\alpha_j u^2},$$

where $w_j = \frac{1}{\Gamma(\gamma)T} \exp(\gamma j/T - e^{j/T})$, and $\alpha_j = e^{j/T}/c^2$. Now, consider a Fourier transform of the target function $F_A(s) = e^{\gamma s - Ae^s}$ in $K_{\gamma,c}(u)$:

$$\begin{aligned} \widehat{F}_A(\xi) &= \int_{-\infty}^\infty e^{\gamma s - Ae^s} e^{-i\xi s} ds \\ &= \int_{-\infty}^\infty e^{(\gamma - i\xi)s - Ae^s} ds \\ &= \int_0^\infty t^{\gamma - i\xi - 1} e^{-At} dt \quad \text{(by taking } t = e^s\text{)} \\ &= A^{-\gamma + i\xi} \Gamma(\gamma - i\xi). \end{aligned}$$

Then,

$$\begin{aligned} I_T(u) &= \frac{1}{\Gamma(\gamma)T} \sum_{j\in\mathbb{Z}} e^{\gamma j/T - e^{j/T}} e^{-e^{j/T}u^2/c^2} \\ &= \frac{1}{\Gamma(\gamma)T} \sum_{j\in\mathbb{Z}} F_A(j/T) \\ &= \frac{1}{2\pi\Gamma(\gamma)T} \int_{\mathbb{R}} \widehat{F}_A(\xi) \sum_{j\in\mathbb{Z}} e^{i\xi j/T} d\xi \\ &= \frac{1}{\Gamma(\gamma)} \sum_{\ell\in\mathbb{Z}} \widehat{F}_A(2\pi T\ell), \end{aligned}$$

where the last step uses the Poisson sum distribution identity. Note that $\int_{\mathbb{R}} F_A(s)ds = \widehat{F}_A(0) = A^{-\gamma}\Gamma(\gamma)$. Hence, the discretization error is

$$\begin{aligned}
|I_T(u) - K_{\gamma,c}(u)| &= \left| I_T(u) - \int_{\mathbb{R}} F_A(s)ds \right| \\
&= \frac{1}{\Gamma(\gamma)} \left| \sum_{\ell \neq 0} \widehat{F}_A(2\pi T\ell) \right| \\
&= \frac{1}{\Gamma(\gamma)} \left| \sum_{\ell \geq 1} \widehat{F}_A(2\pi T\ell) + \sum_{p \geq 1} \widehat{F}_A(-2\pi Tp) \right| \\
&= \frac{1}{\Gamma(\gamma)} \left| \sum_{\ell \geq 1} A^{-\gamma + 2\pi i T\ell} \Gamma(\gamma - 2\pi i T\ell) + \sum_{p \geq 1} A^{-\gamma - 2\pi i T p} \Gamma(\gamma + 2\pi i T p) \right| \\
&\leq \frac{1}{\Gamma(\gamma)} \left[\sum_{\ell \geq 1} A^{-\gamma} |\Gamma(\gamma - 2\pi i T\ell)| + \sum_{p \geq 1} A^{-\gamma} |\Gamma(\gamma + 2\pi i T p)| \right] \\
&\leq \frac{1}{\Gamma(\gamma)} \left[\sum_{\ell \geq 1} |\Gamma(\gamma - 2\pi i T\ell)| + \sum_{p \geq 1} |\Gamma(\gamma + 2\pi i T p)| \right],
\end{aligned}$$

provided that $A \geq 1$. Observe that $|\gamma \pm 2\pi i T\ell| \geq 2\pi$ for all $\ell \geq 1$. Also, note that $\arg(\gamma \pm 2\pi i T\ell) \in (-\pi/2, \pi/2)$, and hence we apply Stirling's formula for the gamma function (see e.g., Montgomery & Vaughan (2006, Theorem C.1)) to get

$$|\Gamma(\gamma \pm 2\pi i T\ell)| \leq C_\gamma \big(1 + |\ell| T\big)^{\gamma - 1/2} \exp\Big(- \pi^2 |\ell| T \Big).$$

We obtain

$$|I_T(u) - K_{\gamma,c}(u)| \leq \frac{2C_\gamma}{\Gamma(\gamma)} \sum_{\ell \geq 1} \big(1 + \ell T\big)^{\gamma - 1/2} \exp\big(- \pi^2 \ell T \big).$$

Now, observe that $C'_\gamma := \sup_{x \geq 1} (1+x)^{\gamma - 1/2} \exp(-\pi^2 x) < \infty$, hence

$$\begin{aligned}
|I_T(u) - K_{\gamma,c}(u)| &\leq \frac{2C_\gamma C'_\gamma}{\Gamma(\gamma)} \sum_{\ell \geq 1} \exp\left(- \frac{\pi^2}{2} \ell T \right) \\
&= \frac{2C_\gamma C'_\gamma}{\Gamma(\gamma)} \cdot \frac{\exp(-\pi^2 T/2)}{1 - \exp(-\pi^2 T/2)} \\
&\leq C''_\gamma e^{-\pi^2 T/2},
\end{aligned}$$

where $C''_\gamma = 4C_\gamma C'_\gamma / \Gamma(\gamma)$ is a constant depending only on γ.

We derive truncation error similarly. Let $L \in \mathbb{N}$ as a truncation parameter, and consider the truncated approximation

$$I_L(u) = \frac{1}{\Gamma(\gamma) T} \sum_{|j| \leq L} e^{\gamma j/T - e^{j/T}} e^{-e^{j/T} u^2/c^2}.$$

Then, we have

$$\begin{aligned}
|I_T(u) - I_{T,L}(u)| &\leq \frac{1}{\Gamma(\gamma) T} \sum_{|j| > L} e^{\gamma j/T - e^{j/T}} e^{-e^{j/T} u^2/c^2} \\
&\leq \frac{1}{\Gamma(\gamma) T} \sum_{|j| > L} e^{\gamma j/T - e^{j/T}},
\end{aligned}$$

where the last inequality follows from $e^{-e^{j/T} u^2/c^2} \leq 1$.

We bound the two tails separately. For the positive tail, define

$$C_{\gamma,+} := \sup_{x\geq 0} \exp\left((\gamma+1)x - e^x\right) < \infty.$$

Then, for $j \geq 1$,

$$e^{\gamma j/T - e^{j/T}} = e^{-j/T} \exp\left((\gamma+1)j/T - e^{j/T}\right) \leq C_{\gamma,+} e^{-j/T}.$$

Hence

$$\begin{aligned}\frac{1}{T}\sum_{j>L} e^{\gamma j/T - e^{j/T}} &\leq \frac{C_{\gamma,+}}{T}\sum_{j>L} e^{-j/T} \\ &= \frac{C_{\gamma,+}}{T}\cdot\frac{e^{-(L+1)/T}}{1-e^{-1/T}} \\ &\leq C'_{\gamma,+} e^{-L/T},\end{aligned}$$

where $C'_{\gamma,+}$ depends only on γ, since $\inf_{T\geq 1} T(1-e^{-1/T}) > 0$.

For the negative tail, writing $j = -m$, we get

$$\begin{aligned}\frac{1}{T}\sum_{j<-L} e^{\gamma j/T - e^{j/T}} &= \frac{1}{T}\sum_{m>L} e^{-\gamma m/T - e^{-m/T}} \\ &\leq \frac{1}{T}\sum_{m>L} e^{-\gamma m/T} \\ &= \frac{1}{T}\cdot\frac{e^{-\gamma(L+1)/T}}{1-e^{-\gamma/T}} \\ &\leq C'_{\gamma,-} e^{-\gamma L/T},\end{aligned}$$

where $C'_{\gamma,-}$ depends only on γ, since $\inf_{T\geq 1} T(1-e^{-\gamma/T}) > 0$.

Therefore, if we put $\eta_\gamma = \min\{1,\gamma\}$, then there exists a constant $C_{2,\gamma} > 0$ depending only on γ such that

$$\sup_{u\in\mathbb{R}} |I_T(u) - I_{T,L}(u)| \leq C_{2,\gamma} e^{-\eta_\gamma L/T}.$$

Combining this with the discretization estimate gives

$$\sup_{u\in\mathbb{R}} |K_{\gamma,c}(u) - I_{T,L}(u)| \leq C_{1,\gamma} e^{-\pi^2 T/2} + C_{2,\gamma} e^{-\eta_\gamma L/T}, \tag{15}$$

for some constant $C_{1,\gamma} > 0$ depending only on γ.

Now we choose T and L so that the two terms in (15) are sufficiently small. Given $0 < \delta < 1$, it suffices to take

$$T \geq \max\left\{1, \frac{2}{\pi^2}\log\left(\frac{3C_{1,\gamma}}{\delta}\right)\right\},$$

and

$$L \geq \frac{T}{\eta_\gamma}\log\left(\frac{3C_{2,\gamma}}{\delta}\right).$$

With these choices,

$$\sup_{u\in\mathbb{R}} |K_{\gamma,c}(u) - I_{T,L}(u)| \leq \frac{2\delta}{3}.$$

It remains to convert the finite Gaussian mixture $I_{T,L}$ into a sum of piecewise separable kernels. Recall that

$$I_{T,L}(|x-y|) = \sum_{|j|\le L} w_j e^{-\alpha_j (x-y)^2},$$

where

$$w_j = \frac{1}{\Gamma(\gamma)T}\exp(\gamma j/T - e^{j/T}), \qquad \alpha_j = \frac{e^{j/T}}{c^2}.$$

By Appendix C.3.2, and by applying the lemma to the rescaled variables $\sqrt{\alpha_j}x$ and $\sqrt{\alpha_j}y$, for every $\varepsilon > 0$ there exists a sum $\widetilde{G}_j$ of $R_G(\varepsilon)$ real piecewise separable kernels such that

$$\sup_{x,y\in\mathbb{R}} \left| e^{-\alpha_j (x-y)^2} - \widetilde{G}_j(x,y) \right| \le \varepsilon,$$

where, for every fixed $\delta_0 \in (0,1)$ and $0 < \varepsilon \le \delta_0$,

$$R_G(\varepsilon) \le C_G(\delta_0) \log\left(\frac{1}{\varepsilon}\right).$$

The constant $C_G(\delta_0)$ is independent of α_j.

Define

$$\widetilde{K}_{\gamma,c}(x,y) = \sum_{|j|\le L} w_j \widetilde{G}_j(x,y).$$

Since nonnegative scalar multiples and finite sums of real piecewise separable kernels are again real piecewise separable, $\widetilde{K}_{\gamma,c}$ is a sum of at most $(2L+1)R_G(\varepsilon)$ real piecewise separable kernels.

We now bound the approximation error. First note that

$$\sum_{|j|\le L} w_j \le \sum_{j\in\mathbb{Z}} w_j = I_T(0) \le 1 + C_{1,\gamma} e^{-\pi^2 T/2} \le B_\gamma$$

for some constant $B_\gamma > 0$ depending only on γ. Hence

$$\begin{aligned}
&\sup_{x,y\in\mathbb{R}} \left| K_{\gamma,c}(|x-y|) - \widetilde{K}_{\gamma,c}(x,y) \right| \\
&\le \sup_{x,y\in\mathbb{R}} |K_{\gamma,c}(|x-y|) - I_{T,L}(|x-y|)| + \sum_{|j|\le L} w_j \sup_{x,y\in\mathbb{R}} \left| e^{-\alpha_j (x-y)^2} - \widetilde{G}_j(x,y) \right| \\
&\le \frac{2\delta}{3} + B_\gamma \varepsilon.
\end{aligned}$$

Taking

$$\varepsilon = \frac{\delta}{3B_\gamma},$$

we obtain

$$\sup_{x,y\in\mathbb{R}} \left| K_{\gamma,c}(|x-y|) - \widetilde{K}_{\gamma,c}(x,y) \right| \le \delta.$$

Finally, we estimate the number of real piecewise separable kernels. With the above choices,

$$T = O_\gamma\left(\log\frac{1}{\delta}\right), \qquad L = O_\gamma\left(\log^2\frac{1}{\delta}\right).$$

Moreover,

$$R_G(\varepsilon) \le C_G(\delta_0) \log\left(\frac{3B_\gamma}{\delta}\right) = O_{\gamma,\delta_0}\left(\log\frac{1}{\delta}\right).$$

Therefore the total number R of real piecewise separable kernels satisfies

$$\begin{aligned} R &\le (2L+1)R_G(\varepsilon) \\ &\le C_{\gamma,\delta_0}\left(\log\frac{1}{\delta}\right)^3, \end{aligned}$$

for all $0 < \delta \le \delta_0 < 1$, where C_{γ,δ_0} depends only on γ and δ_0. This completes the proof.

C.4. Proof of Lemma 4.1

For $\Delta^B_{n,\alpha}(\mathbf{X},\mathbf{Y};\tilde{k})$ to be level α, it suffices to show that $B+1$ random variables $\widehat{\mathrm{MMD}}^2(\mathbf{Z};\tilde{k}), \widehat{\mathrm{MMD}}^2(\mathbf{Z}_{\pi_1};\tilde{k}), \ldots, \widehat{\mathrm{MMD}}^2(\mathbf{Z}_{\pi_B};\tilde{k})$ are exchangeable under the null hypothesis. Note that $\tilde{k}$ is a fixed bivariate function that does not depend on the data or permutations, hence the exchangeability of the above random variables directly follows from the reasoning of Schrab et al. (2023, Appendix E.1) and Albert et al. (2022, Appendix C.1).

Now we analyze the power of the approximated test. First, observe that

$$\begin{aligned} &|\widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};k) - \widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};\tilde{k})| \\ \le &\left|\frac{1}{n(n-1)}\sum_{i\ne j}(k(X_i,X_j) - \tilde{k}(X_i,X_j))\right| + \left|\frac{1}{m(m-1)}\sum_{i\ne j}(k(Y_i,Y_j) - \tilde{k}(Y_i,Y_j))\right| \\ &+\left|\frac{2}{nm}\sum_{i,j}(k(X_i,Y_j) - \tilde{k}(X_i,Y_j))\right| \\ \le &\frac{1}{n(n-1)}\sum_{i\ne j}\left|k(X_i,X_j) - \tilde{k}(X_i,X_j)\right| + \frac{1}{m(m-1)}\sum_{i\ne j}\left|k(Y_i,Y_j) - \tilde{k}(Y_i,Y_j)\right| \\ &+\frac{2}{nm}\sum_{i,j}\left|k(X_i,Y_j) - \tilde{k}(X_i,Y_j)\right| \\ \le &4\delta_n. \end{aligned}$$

Also, the similar bound holds for the permutation quantile. Denoting the empirical estimator with kernel k as $T_k(\mathbf{Z})$, note that for all permutation π, we have $|T_k(\mathbf{Z}_\pi) - T_{\tilde{k}}(\mathbf{Z}_\pi)| \le 4\delta_n$. Observe that $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of $T_k(\mathbf{Z})$, which is the order statistic of $T_k(\mathbf{Z})$. Note that for any $r \in \{1,\ldots,(n+m)!\}$, the r-th order statistic of $T_k(\mathbf{Z})$ and the r-th order statistic of $T_{\tilde{k}}(\mathbf{Z})$ differ by at most $4\delta_n$, which implies that $|q_{1-\alpha}(k) - q_{1-\alpha}(\tilde{k})| \le 4\delta$.

We now compare the rejection events. From the bounds

$$|T_k - T_{\tilde{k}}| \le 4\delta, \quad |q_{1-\alpha,B}(k) - q_{1-\alpha,B}(\tilde{k})| \le 4\delta,$$

we have the event inclusions

$$\{T_k - 8\delta > q_{1-\alpha,B}(k)\} \subseteq \{T_{\tilde{k}} > q_{1-\alpha,B}(\tilde{k})\} \subseteq \{T_k + 8\delta > q_{1-\alpha,B}(k)\}.$$

Taking probabilities under H_1 yields

$$\mathbb{P}_{H_1}\Big(T_k - 8\delta > q_{1-\alpha,B}(k)\Big) \le \mathbb{P}_{H_1}\Big(\Delta^B_\alpha(\mathbf{X},\mathbf{Y};\tilde{k}) = 1\Big) \le \mathbb{P}_{H_1}\Big(T_k + 8\delta > q_{1-\alpha,B}(k)\Big).$$

C.5. Proof of Theorem 4.2

We follow the proof of Schrab et al. (2023). For simplicity, we write $T_k = \widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};k)$ and $T_{\tilde{k}} = \widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y};\tilde{k})$. Also, since the four one-dimensional kernels are translation invariant, note that such kernels k_i admit the function K_i which satisfies $k_i(x,y) = K_i(x-y)$.

Note that we are interested in controlling the type II error as β, i.e.,

$$\mathbb{P}_{X,Y,\pi}\Big(T_{\tilde{k}} \le q_{1-\alpha,B}(\tilde{k})\Big) \le \beta.$$

From Appendix C.4, we have the event inclusion

$$\big\{T_{\tilde{k}} \le q_{1-\alpha,B}(\tilde{k})\big\} \subseteq \big\{T_k \le q_{1-\alpha,B}(k) + 8\delta\big\},$$

and it suffices to show that

$$\mathbb{P}_{X,Y,\pi}\Big(T_k \le q_{1-\alpha,B}(k) + 8\delta\Big) \le \beta.$$

By applying Schrab et al. (2023, Lemma 2), it suffices to show that

$$\mathbb{P}_{X,Y,\pi}\left(\mathrm{MMD}^2(P,Q;k_\lambda) \ge \sqrt{\frac{2}{\beta}\mathrm{Var}_{X,Y}(T_k)} + q_{1-\alpha,B}(k) + 8\delta\right) \ge 1 - \frac{\beta}{2}. \tag{16}$$

With Schrab et al. (2023, Proposition 4) in place, the above probability inequality holds if

$$\mathrm{MMD}^2(P,Q;k_\lambda) \ge \sqrt{\frac{2}{\beta}\mathrm{Var}_{X,Y}(T_k)} + C_2'(M,d)\sqrt{\kappa_2}\sqrt{\frac{2}{\beta}}\frac{\log(1/\alpha)}{(n+m)\sqrt{\lambda_1\cdots\lambda_d}} + 8\delta \tag{17}$$

satisfies for some constant $C_2'(M,d) > 0$, provided that $\alpha \in (0, e^{-1}),\ \beta \in (0, 1-\alpha),\ B \ge 3\alpha^{-2}(\log(8/\beta) + \alpha(1-\alpha))$, and $\|p\|_\infty, \|q\|_\infty \le M$, where κ_2 is defined as $\kappa_2 = \prod_{i=1}^d \int K_i(u)^2 du$. Note that each $\int K_i(u)^2 du$ is finite:

1. For the Laplace kernel, we have $\int K_i(u)^2 du = \int e^{-2|u|} du = \frac{1}{4}$.
2. For the Matern kernel, we have $\int K_i(u)^2 du = \sqrt{\frac{\pi}{p+1/2}}\frac{\Gamma(p+1)^2\Gamma(2p+3/2)}{\Gamma(p+1/2)^2\Gamma(2p+2)}$.
3. For the Gaussian kernel, we have $\int K_i(u)^2 du = \frac{1}{\pi}\int e^{-2u^2} du = \frac{1}{\sqrt{2\pi}}$.
4. For the inverse multiquadric kernel, we have $\int K_i(u)^2 du = c\int (1+t^2)^{-2\gamma} dt = c\sqrt{\pi}\frac{\Gamma(2\gamma-1/2)}{\Gamma(2\gamma)} < \infty$, since $\gamma > 1/2$.

Here, note that the kernels are chosen a priori, and thus κ_2 can be seen as a universal constant and we do not explicitly track the dependence on κ_2 in the subsequent analysis. Hence, we write $C_2(M,d) = C_2'(M,d)\sqrt{\kappa_2}$, which is a constant depending only on M and d.

We note that the variance bound is provided in Schrab et al. (2023, Proposition 3):

$$\mathrm{Var}_{X,Y}(T_k) \le C_1(M,d)\left(\frac{\|\psi * \varphi_\lambda\|_2^2}{n+m} + \frac{1}{(n+m)^2\lambda_1\cdots\lambda_d}\right),$$

for some constant $C_1(M,d) > 0$, and the fact (Schrab et al., 2023, Appendix E.5) that $\mathrm{MMD}^2(P,Q;k_\lambda)$ can be written as

$$\mathrm{MMD}^2(P,Q;k_\lambda) = \frac{1}{2}\big(\|\psi\|_2^2 + \|\psi * \varphi_\lambda\|_2^2 - \|\psi - \psi * \varphi_\lambda\|_2^2\big),$$

where $\psi = p - q$ and $\varphi_\lambda(x-y) = k_\lambda(x,y)$ (which is valid since the components in k_λ are chosen in Lemma C.1, and thus k_λ is translation invariant). With these in place, the sufficient condition (17) is implied by

$$\|\psi\|_2^2 \ge \|\psi - \psi * \varphi_\lambda\|_2^2 + C_3(M,d)\frac{\log(1/\alpha)}{\beta(n+m)\sqrt{\lambda_1\cdots\lambda_d}} + 16\delta, \tag{18}$$

for some constant $C_3(M,d) > 0$, provided that $\alpha \in (0, e^{-1})$. Since there exists some constant $S \in (0,1)$ (Schrab et al., 2023, Appendix E.6) satisfying

$$\|\psi - \psi * \varphi_\lambda\|_2^2 - S\|\psi\|_2^2 \le C_4(d,s,R)\sum_{i=1}^d \lambda_i^{2s},$$

a sufficient condition (18) is

$$\|\psi\|_2^2 \geq C_5(M,d,s,R,\alpha,\beta)\left(\sum_{i=1}^d \lambda_i^{2s} + \frac{1}{(n+m)\sqrt{\lambda_1\cdots\lambda_d}}\right) + \frac{16}{1-S}\delta.$$

With the choice of $\lambda_i^* = (n+m)^{-\frac{2}{4s+d}}$ for $i = 1,\ldots,d$, the sufficient condition becomes

$$\|\psi\|_2^2 \geq C_6(M,d,s,R,\alpha,\beta)(n+m)^{-\frac{4s}{4s+d}} + \frac{16}{1-S}\delta,$$

which suffices if

$$\|\psi\|_2 \geq C_7(M,d,s,R,\alpha,\beta)(n+m)^{-\frac{2s}{4s+d}} + \frac{4}{\sqrt{1-S}}\sqrt{\delta},$$

since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. This implies that the uniform separation rate of the approximated MMD test is

$$\rho_n(\tilde{\Delta}_{n,\alpha}^{B,\lambda^*};\beta,\mathcal{C}) \leq C_7(M,d,s,R,\alpha,\beta)(n+m)^{-\frac{2s}{4s+d}} + \frac{4}{\sqrt{1-S}}\sqrt{\delta}.$$

Therefore, if there exists some constant $C > 0$ such that δ satisfies

$$\delta \leq C(n+m)^{-\frac{4s}{4s+d}},$$

then the approximated MMD test is minimax rate optimal.

Now, note that if each k_i admits (R_i, δ_i)-APS approximation where δ_i is uniformly upper bounded by a constant δ_0, then by Theorem 3.3, the product kernel k_λ admits an (R,δ)-AOS approximation with

$$R = \prod_{i=1}^d R_i, \quad \delta = \frac{(U+\delta_0)^{d-1}d}{\lambda_1\cdots\lambda_d}\delta_{\max},$$

where $\delta_{\max} = \max_{i=1,\ldots,d}\delta_i$, and $U = \max_{i=1,\ldots,d} U_i$. With the choice of $\lambda_i^* = (n+m)^{-\frac{2}{4s+d}}$, it suffices to have

$$\frac{(U+\delta_0)^{1-d}}{d}\delta_{\max} \leq C(n+m)^{-\frac{4s+2d}{4s+d}},$$

for some constant $C > 0$. Since we are only considering four kernels in Lemma C.1, U is less than or equal to 1. Note that d is fixed. Now, take $\delta_0 = 0.5$. This allow us to write the upper bound of $\frac{(U+\delta_0)^{1-d}}{d}$ as a universal constant $1 \geq C' > 0$. Take $\delta_i = C'(n+m)^{-\frac{4s+2d}{4s+d}}$ for $i = 1,\ldots,d$, and observe that this is always less than or equal to 0.5 for all $n, m \geq 1$. Then, for all $i = 1,\ldots,d$, we have $\delta_i \leq \delta_0$. Then, note that the sufficient condition for minimax optimality

$$\delta_{\max} \leq C'(n+m)^{-\frac{4s+2d}{4s+d}}$$

is satisfied with universal constant C'. We display the number of real piecewise separable kernels in the approximation for each kernel in Theorem 4.2 as follows:

1. For Laplace kernel k_i, $R_i = 1$ is sufficient.
2. For Matérn kernel k_i with half-integer $\nu = p + 1/2$ for some $p \in \mathbb{N}$, $R_i = p + 1$ is sufficient.
3. For Gaussian kernel k_i,

$$R_i = \left\lceil C_G \log\left(\frac{1}{\min\left\{0.5, C'(n+m)^{-\frac{4s}{4s+d}}\right\}}\right)\right\rceil$$

is sufficient, where $C_G > 0$ is a fixed constant. Since $C' \leq 1$, we have $\min\{0.5, C'(n+m)^{-\frac{4s}{4s+d}}\} = C'(n+m)^{-\frac{4s}{4s+d}}$. Therefore, we can write R_i as

$$R_i = \left\lceil C_G \frac{4s+2d}{4s+d}\log(n+m)\right\rceil.$$

4. For inverse multiquadric kernel k_i,

$$R_i = \left\lceil C_{\mathrm{IMQ},\gamma} \log^3 \left(\frac{1}{\min\left\{0.5, C'(n+m)^{-\frac{4s}{4s+d}}\right\}} \right) \right\rceil$$

is sufficient, where $C_{\mathrm{IMQ},\gamma} > 0$ is a fixed constant depending only on γ. Similar to the Gaussian kernel case, we can write R_i as

$$R_i = \left\lceil C_{\mathrm{IMQ},\gamma} \left(\frac{4s+2d}{4s+d} \right)^3 \log^3 \left(n+m\right) \right\rceil.$$

Therefore, if the kernel k_λ is a product of l_k, g_k, m_k, and h_k Laplace, Gaussian, Matérn with half-integer smoothness, and inverse multiquadric kernels respectively, then the total number of real piecewise separable kernels in the approximation satisfies

$$\begin{aligned} R = \prod_{i=1}^{d} R_i &= (p+1)^{m_k} \left\lceil C_G \frac{4s+2d}{4s+d} \log\left(n+m\right) \right\rceil^{g_k} \left\lceil C_{\mathrm{IMQ},\gamma} \left(\frac{4s+2d}{4s+d} \right)^3 \log^3\left(n+m\right) \right\rceil^{h_k} \\ &\leq C(p,\gamma,s,d)(p+1)^{m_k} \log^{3h_k+g_k}(n+m), \end{aligned}$$

where $C(p,\gamma,s,d) > 0$ is a constant depending only on p, γ, s, and d. This completes the proof.

C.5.1. MMD Optimality

Another common framework (Domingo-Enrich et al., 2023; Kim & Schrab, 2026) for kernel two sample testing is to use the MMD itself as the distance between two distributions, i.e., $d(P,Q) = \mathrm{MMD}(P,Q;k)$, where k is a characteristic kernel. As shown in (Kim & Schrab, 2026), the minimax separation rate is $\epsilon_n^* \asymp n^{-1/2}$. We show that the approximated MMD test with the product of kernels in Lemma C.1 again achieves the minimax separation rate ϵ_n^* up to a constant factor, within near-linear time complexity.

Corollary C.4. *Consider the two sample testing problem defined in Equation (1) with $\mathcal{C}_1 = \mathcal{P}(\mathbb{R}^d)$. Assume that $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, and $B \geq 3\alpha^{-2}(\log(8/\beta) + \alpha(1-\alpha))$. Consider the product kernel k_λ defined in Theorem 3.3 with each k_i being Laplace, Matérn, Gaussian, or inverse multiquadric kernel in Lemma C.1. For any fixed $\lambda \in (0,\infty)^d$, there exists a $\tilde{k}_{\lambda^*}$ which is (R,δ)-AOS approximation of k_λ with*

$$R = (p+1)^{m_k} \log^{3h_k+g_k} N, \quad \delta \leq C(n+m)^{-1/2},$$

where $C > 0$ is a constant, and l_k, m_k, g_k, h_k denote the number of Laplace, Matérn, Gaussian, and inverse multiquadric kernels producted in the kernel k_λ, respectively. The corresponding approximated test $\tilde{\Delta}_\alpha^{B,\lambda}$ with such $\tilde{k}_\lambda$ is minimax rate optimal, i.e.,

$$\rho_n(\tilde{\Delta}_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1) \leq C(\alpha,\beta,p,\gamma)\epsilon_n^*(\beta,\mathcal{C}_1).$$

Proof. Given fixed $k_1,\ldots,k_d$ and bandwidth $\lambda_1,\ldots,\lambda_d > 0$, let us denote

$$K = \sup_{x,y\in\mathbb{R}^d} |k_\lambda(x,y)| = \frac{1}{\lambda_1\cdots\lambda_d} \sup_{x,y\in\mathbb{R}} |k(x,y)|.$$

We first point out that unlike L_2-separation, the kernel and bandwidth should be fixed for the MMD separation, otherwise the the metric induced by the kernel will also change. Hence, we treat K as a fixed constant. We start from Equation (16):

$$\mathbb{P}_{X,Y,\pi}\left(\mathrm{MMD}^2(P,Q;k_\lambda) \geq \sqrt{\frac{2}{\beta}\mathrm{Var}_{X,Y}(T_k)} + q_{1-\alpha,B}(k) + 8\delta \right) \geq 1 - \frac{\beta}{2},$$

is sufficient to control the type II error by β. By Schrab et al. (2023, Proposition 4), the quantile $q_{1-\alpha,B}(k)$ has a high-probability upper bound:

$$\mathbb{P}_{X,Y,\pi}\big(q_{1-\alpha,B}(k) \leq q_{1-\alpha/2,\infty}(k)\big) \geq 1 - \frac{\beta}{2},$$

if $B \geq \frac{3}{\alpha^2}\big(\log\big(\frac{4}{\beta}\big) + \alpha(1-\alpha)\big)$, where $q_{1-\alpha/2,\infty}(k)$ is the $1-\alpha/2$ quantile of the empirical distribution of T_k considering all possible permutations, i.e.,

$$q_{1-\alpha/2,\infty}(k) = \inf\left\{t \in \mathbb{R} : \frac{1}{(n+m)!}\sum_{\pi} \mathbf{1}\big(T_k(\mathbf{Z}_\pi) \leq t\big) \geq 1-\alpha/2\right\}.$$

Therefore, it suffices to show the deterministic inequality

$$\mathrm{MMD}^2(P,Q;k_\lambda) \geq \sqrt{\frac{2}{\beta}\mathrm{Var}_{X,Y}(T_k)} + q_{1-\alpha/2,\infty}(k) + 8\delta. \tag{19}$$

Now, by following similar argument in Choi & Kim (2024, Page 36) or Kim et al. (2022, Appendix C) , we obtain the upper bound for the quantile $q_{1-\alpha/2,\infty}(k)$:

$$q_{1-\alpha/2,\infty}(k) \leq C(\alpha)KN^{-1},$$

where $C(\alpha) > 0$ is a constant depending only on α. Also, from Choi & Kim (2024, Equation 55), we have the variance bound

$$\sqrt{\frac{2}{\beta}\mathrm{Var}_{X,Y}(T_k)} \leq C(\beta)\Big(\sqrt{K}N^{-1/2}\mathrm{MMD}(P,Q;k_\lambda) + KN^{-1}\Big),$$

where $C(\beta) > 0$ is a constant depending only on β. Summarizing the above bounds, if we have

$$\begin{aligned}
&\mathrm{MMD}^2(P,Q;k_\lambda) \geq 3(C(\alpha)+C(\beta))KN^{-1},\\
&\mathrm{MMD}^2(P,Q;k_\lambda) \geq 3C(\beta)\sqrt{K}N^{-1/2}\mathrm{MMD}(P,Q;k_\lambda),\\
&\mathrm{MMD}^2(P,Q;k_\lambda) \geq 24\delta,
\end{aligned}$$

simultaneously, then Equation (19) holds, by the fact that $3a, 3b, 3c \leq d$ implies $a+b+c \leq d$. Note that the first two inequalities are satisfied by

$$\mathrm{MMD}(P,Q;k_\lambda) \geq C(\alpha,\beta)KN^{-1/2},$$

for some constant $C(\alpha,\beta) > 0$ depending only on α and β. Now suppose that $\delta \leq C_\delta N^{-1}$ for some constant $C_\delta > 0$. Then, the third inequality is implied by

$$\mathrm{MMD}(P,Q;k_\lambda) \geq C_\delta N^{-1/2},$$

which implies that the uniform separation rate of the approximated MMD test is

$$\rho_n(\tilde{\Delta}^{B,\lambda}_{n,\alpha};\beta,\mathcal{C}) \leq C(\alpha,\beta,C_\delta)N^{-1/2},$$

where we take $C(\alpha,\beta,C_\delta) = \max\{C(\alpha,\beta)K, C_\delta\}$.

Therefore, if there exists some constant $C > 0$ such that δ satisfies

$$\delta \leq CN^{-1},$$

then the approximated MMD test is minimax optimal in terms of the MMD separation rate. Note that if each k_i admits (R_i,δ_i)-APS approximation where δ_i is uniformly upper bounded by a constant δ_0, then by Theorem 3.3, the product kernel k_λ admits an (R,δ)-AOS approximation with

$$R = \prod_{i=1}^{d} R_i, \quad \delta = \frac{(U+\delta_0)^{d-1}d}{\lambda_1\cdots\lambda_d}\delta_{\max},$$

where $\delta_{\max} = \max_{i=1,\dots,d}\delta_i$, and $U = \max_{i=1,\dots,d}U_i$. Since $\lambda_1\cdots\lambda_d$ is a fixed constant, it suffices to have

$$\frac{(U+\delta_0)^{1-d}}{d}\delta_{\max} \leq CN^{-1},$$

for some constant $C > 0$. Now, take $\delta_0 = 0.5$. Since we are only considering four kernels in Lemma C.1, U is less than or equal to 1. Note that d is fixed. Then, there exists a constant $1 > C' > 0$ such that $C' > \frac{(U+\delta_0)^{1-d}}{d}$. Now, for each $i = 1, \ldots, d$, take $\delta_i = C'N^{-1}$. Then, note that the sufficient condition for minimax optimality

$$\delta_{\max} \leq C'N^{-1}$$

is satisfied with constant C', and δ_i is uniformly upper bounded by 0.5 for all $i = 1, \ldots, d$. From the analysis in the proof of Theorem 4.2, the number of real piecewise separable kernels in approximating k_λ satisfies

$$\begin{aligned} R = \prod_{i=1}^{d} R_i &= (p+1)^{m_k} \left\lceil C_G \log\left(N\right) \right\rceil^{g_k} \left\lceil C_{\mathrm{IMQ},\gamma} \log^3\left(N\right) \right\rceil^{h_k} \\ &\leq C(p,\gamma)(p+1)^{m_k} \log^{3h_k+g_k}(N), \end{aligned}$$

where $C(p,\gamma) > 0$ is a constant depending only on p and γ. This completes the proof.

C.6. Proof of Lemma 4.4

C.6.1. Lower bound analysis

We follow the same construction as in the proof of Appendix B, and we refer to Schrab et al. (2023); Li & Yuan (2024); Albert et al. (2022) for the choice of $p_0^{(n)}$ and $\{\psi_{n,k}\}_{k \in \mathcal{K}_n}$. Instead of mere $[0,1]^d$-supported densities, we consider the following baseline density $p_0^{(n)}$:

$$p_0^{(n)}(x) := \prod_{i=1}^{d} h \mathbf{1}_{[0,1/h]}(x_i), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

where $h \in (0, \infty)$. Consider a baseline bump function $\psi_0 : \mathbb{R} \to \mathbb{R}$ defined as

$$\psi_0(u) := \exp\left(-\frac{1}{1-(4u-1)^2}\right)\mathbf{1}_{(0,1/2)}(u) - \exp\left(-\frac{1}{1-(4u-3)^2}\right)\mathbf{1}_{(1/2,1)}(u), \quad u \in \mathbb{R}.$$

We define d-dimensional bump functions $\psi_n : \mathbb{R}^d \to \mathbb{R}$ as

$$\psi_n(x) := \prod_{i=1}^{d} \psi_0(x_i), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

and its localized version $\psi_{n,k} : \mathbb{R}^d \to \mathbb{R}$ as

$$\psi_{n,k}(x) := \psi_n(hb_n x - k), \quad k \in \mathcal{K}_n.$$

Intuitively, we consider $p_0^{(n)}(x)$ as a uniform density supported on $[0, 1/h]^d$, and $\psi_{n,k}(x)$ as a localized bump function supported on the hypercube $\prod_{i=1}^{d}[k_i 1/(hb_n), (k_i+1)/(hb_n)]$ for each $k = (k_1, \ldots, k_d) \in \mathcal{K}_n$. For notational purpose, we write $h_n = hb_n$.

Risk control via TV distance We first consider Equation (6):

$$b_n^d n^2 \omega_n^4 \max_{k \in \mathcal{K}_n} v_{n,k}^2 \leq c(\alpha, \beta, \eta).$$

Note that for each $k \in \mathcal{K}_n$, we have

$$\begin{aligned} v_{n,k} &= \int_{\mathbb{R}^d} \frac{\psi_{n,k}(x)^2}{p_0^{(n)}(x)} dx \\ &= h^{-d} \int_{\mathbb{R}^d} \psi_{n,k}(x)^2 dx \\ &= h^{-d} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left(\prod_{i=1}^d \left\{ \psi_0\big(h_n x_i - (k_i - 1)\big) \right\}^2 \right) dx_1 \cdots dx_d \\ &= h^{-d} \left[\int_{\mathbb{R}} |\psi_0(t)|^2 h_n^{-1} dt \right]^d, \quad \text{(by taking } \ h_n x_i = t) \\ &= h^{-d} h_n^{-d} \|\psi_0\|_2^{2d}. \end{aligned}$$

Therefore, the above condition is satisfied if we have

$$b_n^d h^{-2d} n^2 \omega_n^4 h_n^{-2d} \|\psi_0\|_2^{4d} \leq c(\alpha, \beta, \eta),$$

which is equivalent to

$$b_n^{-d} h^{-4d} n^2 \omega_n^4 \|\psi_0\|_2^{4d} \leq c(\alpha, \beta, \eta). \tag{20}$$

Validity of the perturbation densities Next, we check the validity of the perturbation densities $q_{a_n}^{(n)}$ defined as

$$q_{a_n}^{(n)}(x) = p_0^{(n)}(x) + \omega_n \sum_{k \in \mathcal{K}_n} a_{n,k} \psi_{n,k}(x),$$

i.e., we check whether $q_{a_n}^{(n)}$ are non-negative and its sup norm is bounded by M. We refer to Albert et al. (2022, Appendix C.19). Note that

$$\begin{aligned} \left| \sum_{k \in \mathcal{K}_n} a_{n,k} \psi_{n,k}(x) \right| &= \left| \sum_{k \in \mathcal{K}_n} a_{n,k} \prod_{i=1}^d \left\{ \psi_0\big(h_n x_i - k_i\big) \right\}(x) \right| \\ &\leq \left(\sup_x |\psi_0(x)| \right)^d \\ &= e^{-d}. \end{aligned}$$

Therefore, for $x \in [0,1]^d$,

$$\begin{aligned} q_{a_n}^{(n)}(x) &\geq h^d - \left| \omega_n \sum_{k \in \mathcal{K}_n} a_{n,k} \psi_{n,k}(x) \right| \\ &\geq h^d - \omega_n e^{-d}, \end{aligned}$$

and hence,

$$q_{a_n}^{(n)}(x) \geq 0$$

if we have

$$\omega_n \leq h^d e^d.$$

On the other hand, for $x \in [0,1]^d$, we have

$$\begin{aligned} q_{a_n}^{(n)}(x) &\leq h_d + \left| \omega_n \sum_{k \in \mathcal{K}_n} a_{n,k} \psi_{n,k}(x) \right| \\ &\leq h_d + \omega_n e^{-d}, \end{aligned}$$

and hence,

$$\|q_{a_n}^{(n)}\|_\infty \le M_d$$

if we have

$$\omega_n \le e^d(M_d - h^d).$$

Combining the above two conditions, it suffices to have

$$\omega_n \le e^d \min\{h^d, M_d - h^d\}. \tag{21}$$

Inclusion in the Sobolev ball Finally, we check whether $p_0^{(n)} - q_{a_n}^{(n)}$ are included in the Sobolev ball $\mathcal{S}_d^s(\mathcal{R}_d)$. We follow the analysis in Albert et al. (2022, Appendix C.20). Note that the Fourier transform of $q_{a_n}^{(n)} - p_0^{(n)}$ is given by

$$\widehat{q}_{a_n}^{(n)}(\xi) - \widehat{p}_0^{(n)}(\xi) = \omega_n \sum_{k\in\mathcal{K}_n} a_{n,k}\widehat{\psi}_{n,k}(\xi),$$

which implies that

$$\begin{aligned}
\left|\widehat{q}_{a_n}^{(n)}(\xi) - \widehat{p}_0^{(n)}(\xi)\right|^2 &= \omega_n^2\left|\sum_{k\in\mathcal{K}_n} a_{n,k}\widehat{\psi}_{n,k}(\xi)\right|^2 \\
&= \omega_n^2 \sum_{k\in\mathcal{K}_n} a_{n,k}^2|\widehat{\psi}_{n,k}(\xi)|^2 + \omega_n^2 \sum_{k\neq k'} a_{n,k}a_{n,k'}\widehat{\psi}_{n,k}(\xi)\overline{\widehat{\psi}_{n,k'}(\xi)} \\
&= \omega_n^2 \sum_{k\in\mathcal{K}_n} |\widehat{\psi}_{n,k}(\xi)|^2 + \omega_n^2 \sum_{k\neq k'} a_{n,k}a_{n,k'}\widehat{\psi}_{n,k}(\xi)\overline{\widehat{\psi}_{n,k'}(\xi)} \\
&=: D(\xi) + C(\xi),
\end{aligned}$$

where $D(\xi)$ is the sum of the diagonal terms, and $C(\xi)$ is the sum of the cross terms. Now, observe that the Sobolev norm of $p_0^{(n)} - q_{a_n}^{(n)}$ can be computed as

$$\int_{\mathbb{R}^d} \|\xi\|^{2s}|\widehat{q}_{a_n}^{(n)}(\xi) - \widehat{p}_0^{(n)}(\xi)|^2 d\xi = \int_{\mathbb{R}^d} \|\xi\|^{2s} D(\xi)d\xi + \int_{\mathbb{R}^d} \|\xi\|^{2s} C(\xi)d\xi.$$

Since

$$\begin{aligned}
\widehat{\psi}_{n,k}(\xi) &= \int_{\mathbb{R}^d} \psi_{n,k}(x)e^{-i\xi^T x}dx \\
&= \int_{\mathbb{R}^d} \prod_{i=1}^d \psi_0(h_n x_i - k_i)e^{-i\xi^T x}dx \\
&= \prod_{i=1}^d \int_{\mathbb{R}} \psi_0(h_n x_i - k_i)e^{-i\xi_i x_i}dx_i \\
&= \prod_{i=1}^d \int_{\mathbb{R}} \psi_0(t)e^{-i\xi_i(t+k_i)/h_n}h_n^{-1}dt \\
&= h_n^{-d}\prod_{i=1}^d \exp(-i\xi_i k_i/h_n)\widehat{\psi}_0(\xi_i/h_n),
\end{aligned}$$

the diagonal part can be computed as

$$\begin{aligned}
D(\xi) &= \omega_n^2 \sum_{k\in\mathcal{K}_n} |\widehat{\psi}_{n,k}(\xi)|^2 \\
&= \omega_n^2 \sum_{k\in\mathcal{K}_n} h_n^{-2d}\prod_{i=1}^d |\widehat{\psi}_0(\xi_i/h_n)|^2 \\
&= \omega_n^2 b_n^d h_n^{-2d}\prod_{i=1}^d |\widehat{\psi}_0(\xi_i/h_n)|^2
\end{aligned}$$

provided that $|\exp(-i\xi_i k_i/b_n)| = 1$ for each $i \in \{1, \dots, d\}$ and $k \in \mathcal{K}_n$. Then,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|\xi\|^{2s} D(\xi) d\xi \\
&= \omega_n^2 b_n^d h_n^{-2d} \int_{\mathbb{R}^d} \|\xi\|^{2s} \prod_{i=1}^d |\widehat{\psi}_0(\xi_i/h_n)|^2 d\xi \\
&= d^{s-1} \omega_n^2 b_n^d h_n^{-2d} \int_{\mathbb{R}^d} \left(\sum_{i=1}^d |\xi_i|^{2s} \right) \prod_{i=1}^d |\widehat{\psi}_0(\xi_i/h_n)|^2 d\xi \quad (\because \|x\|^{2s} \le d^{s-1} \sum_{i=1}^d |x_i|^{2s} \text{ for } s \ge 1) \\
&= d^s \omega_n^2 b_n^d h_n^{-2d} \int_{\mathbb{R}} |u|^{2s} |\widehat{\psi}_0(u/h_n)|^2 du \left(\int_{\mathbb{R}} |\widehat{\psi}_0(u/h_n)|^2 \right)^{d-1} dw \\
&= d^s \omega_n^2 b_n^d h_n^{-2d} \cdot h_n^{2s} \cdot h_n^d \int_{\mathbb{R}} |t|^{2s} |\widehat{\psi}_0(t)|^2 dt \left(\int_{\mathbb{R}} |\widehat{\psi}_0(t)|^2 dt \right)^{d-1} \quad (\text{by taking } u = b_n t) \\
&\le C_s d^s \omega_n^2 b_n^d h_n^{2s-d} (2\pi)^{d-1} \left(\int_{\mathbb{R}} |\psi_0(t)|^2 dt \right)^{d-1}, \quad (\text{by Plancherel's theorem}), \\
&= c(s) d^s \omega_n^2 b_n^d h_n^{2s-d} (2\pi)^d \|\psi_0\|_2^{2d},
\end{aligned} \tag{22}$$

where $C_s > 0$ is a constant satisfying $\int_{\mathbb{R}} |t|^{2s} |\widehat{\psi}_0(t)|^2 dt \le C_s$. Now we analyze the quantity in the Sobolev norm associated with cross term $C(\xi)$. Observe

$$\int_{\mathbb{R}^d} \|\xi\|^{2s} C(\xi) d\xi = \omega_n^2 \sum_{k \ne k'} a_{n,k} a_{n,k'} \int_{\mathbb{R}^d} \|\xi\|^{2s} \widehat{\psi}_{n,k}(\xi) \overline{\widehat{\psi}_{n,k'}(\xi)} d\xi.$$

We provide a high probability upper bound for $\int_{\mathbb{R}^d} \|\xi\|^{2s} C(\xi) d\xi$. Write

$$\begin{aligned}
\int_{\mathbb{R}^d} \|\xi\|^{2s} C(\xi) d\xi &= \omega_n^2 \sum_{k \ne k'} a_{n,k} a_{n,k'} \int_{\mathbb{R}^d} \|\xi\|^{2s} \underbrace{\widehat{\psi}_{n,k}(\xi) \overline{\widehat{\psi}_{n,k'}(\xi)}}_{=: \Psi_{n,k,k'}(\xi)} d\xi \\
&=: \omega_n^2 \sum_{k \prec k'} a_{n,k} a_{n,k'} \int_{\mathbb{R}^d} \|\xi\|^{2s} \big(\Psi_{n,k,k'}(\xi) + \overline{\Psi_{n,k,k'}(\xi)} \big) d\xi \quad (\because \Psi_{n,k,k'}(\xi) = \overline{\Psi_{n,k',k}(\xi)}),
\end{aligned}$$

where $k \prec k'$ implies that k is lexicographically smaller than k'. Let

$$Z_{n,k,k'} := a_{n,k} a_{n,k'} \int_{\mathbb{R}^d} \|\xi\|^{2s} \big(\Psi_{n,k,k'}(\xi) + \overline{\Psi_{n,k,k'}(\xi)} \big) d\xi,$$

and note that $\{Z_{n,k,k'}\}_{k \prec k'}$ are not mutually independent, thus we write the cross term as a quadratic form in independent Rademacher variables and apply the Hanson–Wright inequality (Rudelson & Vershynin, 2013). Let

$$B_{n,k,k'} := \int_{\mathbb{R}^d} \|\xi\|^{2s} \left(\Psi_{n,k,k'}(\xi) + \overline{\Psi_{n,k,k'}(\xi)} \right) d\xi.$$

Define a symmetric matrix M_n by

$$(M_n)_{k,k'} = \begin{cases} B_{n,k,k'}/2, & k \prec k', \\ B_{n,k',k}/2, & k' \prec k, \\ 0, & k = k'. \end{cases}$$

Then

$$\sum_{k \prec k'} a_{n,k} a_{n,k'} B_{n,k,k'} = a_n^\top M_n a_n$$

and $\mathbb{E}[a_n^\top M_n a_n] = 0$. Hence, by the Hanson–Wright inequality (Rudelson & Vershynin, 2013), there exist universal constants $c > 0$ such that, for all $x > 0$,

$$\mathbb{P}\left(\left|\sum_{k \prec k'} a_{n,k} a_{n,k'} B_{n,k,k'}\right| \geq x\right) \leq 2\exp\left[-c\min\left(\frac{x^2}{\|M_n\|_F^2}, \frac{x}{\|M_n\|_{\mathrm{op}}}\right)\right].$$

Let $m_n := |\mathcal{K}_n| = b_n^d$. Suppose that there exists $U(h_n, d, s)$ an upper bound satisfying $\left|Z_{n,k,k'}\right| \leq U(h_n, d, s)$ for each $k \prec k'$. Since $|a_{n,k} a_{n,k'}| = 1$, the bound $|Z_{n,k,k'}| \leq U(h_n, d, s)$ is equivalent to $|B_{n,k,k'}| \leq U(h_n, d, s)$. Hence,

$$\begin{aligned}
\|M_n\|_F^2 &= \sum_{k,k' \in \mathcal{K}_n} (M_n)_{k,k'}^2 \\
&= \frac{1}{2}\sum_{k \prec k'} B_{n,k,k'}^2 \\
&\leq \frac{1}{2}\binom{m_n}{2} U(h_n, d, s)^2 \\
&\leq \frac{1}{4} b_n^{2d} U(h_n, d, s)^2,
\end{aligned}$$

and therefore

$$\|M_n\|_{\mathrm{op}} \leq \|M_n\|_F \leq \frac{1}{2} b_n^d U(h_n, d, s).$$

Consequently, after changing the universal constant $c > 0$ if necessary,

$$\mathbb{P}\left(\left|\sum_{k \prec k'} a_{n,k} a_{n,k'} B_{n,k,k'}\right| \geq x\right) \leq 2\exp\left[-c\min\left(\frac{x^2}{b_n^{2d} U(h_n, d, s)^2}, \frac{x}{b_n^d U(h_n, d, s)}\right)\right].$$

Equivalently, for $L_\eta := \log(2/\eta)$, there exists a universal constant $C > 0$ such that

$$\left|\sum_{k \prec k'} a_{n,k} a_{n,k'} B_{n,k,k'}\right| \leq C b_n^d U(h_n, d, s)\left(\sqrt{L_\eta} + L_\eta\right)$$

with probability at least $1 - \eta$. Therefore, with probability at least $1 - \eta$, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \|\xi\|^{2s} C(\xi) d\xi &\leq \omega_n^2 \left|\sum_{k \prec k'} Z_{n,k,k'}\right| \\
&\leq C\omega_n^2 b_n^d U(h_n, d, s)\left(\sqrt{\log\frac{2}{\eta}} + \log\frac{2}{\eta}\right).
\end{aligned}$$

Now, we analyze the upper bound for $U(h_n, d, s)$. First, observe

$$\begin{aligned}
\Psi_{n,k,k'}(\xi) &= \widehat{\psi}_{n,k}(\xi)\overline{\widehat{\psi}_{n,k'}(\xi)} \\
&= h_n^{-2d}\prod_{i=1}^d \exp(-i\xi_i k_i/h_n)\widehat{\psi}_0(\xi_i/h_n)\prod_{i=1}^d \exp(i\xi_i k_i'/h_n)\overline{\widehat{\psi}_0(\xi_i/h_n)} \\
&= h_n^{-2d}\prod_{i=1}^d \exp(-i\xi_i(k_i - k_i')/h_n)|\widehat{\psi}_0(\xi_i/h_n)|^2,
\end{aligned}$$

which yields that

$$|\Psi_{n,k,k'}(\xi)| = h_n^{-2d}\prod_{i=1}^d |\widehat{\psi}_0(\xi_i/h_n)|^2.$$

Therefore,

$$\begin{aligned}\left|Z_{n,k,k'}\right| &\leq \int_{\mathbb{R}^d} \|\xi\|^{2s}\left(\left|\Psi_{n,k,k'}(\xi)\right| + \left|\overline{\Psi_{n,k,k'}(\xi)}\right|\right)d\xi \\ &= 2\int_{\mathbb{R}^d} \|\xi\|^{2s}\left|\Psi_{n,k,k'}(\xi)\right|d\xi \\ &= 2h_n^{-2d}\int_{\mathbb{R}^d} \|\xi\|^{2s}\prod_{i=1}^{d}|\widehat{\psi}_0(\xi_i/h_n)|^2 d\xi.\end{aligned}$$

Following the same analysis as in Equation (22), we have

$$\int_{\mathbb{R}^d} \|\xi\|^{2s}\prod_{i=1}^{d}|\widehat{\psi}_0(\xi_i/h_n)|^2 d\xi \leq c(s)d^{s+1}h_n^{2s+d}(2\pi)^d\|\psi_0\|_2^{2d},$$

which implies that

$$\begin{aligned}\left|Z_{n,k,k'}\right| &\leq 2c(s)d^s h_n^{2s-d}(2\pi)^d\|\psi_0\|_2^{2d} \\ &=: U(h_n, d, s).\end{aligned}$$

Summarizing the above analysis, we have that with probability at least $1-\eta$,

$$\int_{\mathbb{R}^d} \|\xi\|^{2s}C(\xi)d\xi \leq c(s)d^s\omega_n^2 b_n^d h_n^{2s-d}(2\pi)^d\|\psi_0\|_2^{2d}\left(\sqrt{\log\frac{2}{\eta}} + \log\frac{2}{\eta}\right).$$

Combining the above analysis of the diagonal term and the cross term, we have

$$\begin{aligned}&\int_{\mathbb{R}^d} \|\xi\|^{2s}|\widehat{q}_{a_n}^{(n)}(\xi) - \widehat{p}_0^{(n)}(\xi)|^2 d\xi \\ =& \int_{\mathbb{R}^d} \|\xi\|^{2s}D(\xi)d\xi + \int_{\mathbb{R}^d} \|\xi\|^{2s}C(\xi)d\xi \\ \leq& c(s)d^s\omega_n^2 b_n^d h_n^{2s-d}(2\pi)^d\|\psi_0\|_2^{2d} + c(s)d^s\omega_n^2 b_n^d h_n^{2s-d}(2\pi)^d\|\psi_0\|_2^{2d}\left(\sqrt{\log\frac{2}{\eta}} + \log\frac{2}{\eta}\right) \\ =& c(s,\eta)d^s\omega_n^2 b_n^d h_n^{2s-d}(2\pi)^d\|\psi_0\|_2^{2d},\end{aligned}$$

with probability at least $1-\eta$, where $c(s,\eta) > 0$ is a constant depending on s and η.

Therefore, a sufficient condition for $q_{a_n}^{(n)} \in \mathcal{S}_d^s(\mathcal{R}_d)$ with probability at least $1-\eta$ is

$$\begin{aligned}& c(s,\eta)d^s\omega_n^2 b_n^d h_n^{2s-d}(2\pi)^d\|\psi_0\|_2^{2d} \leq (2\pi)^d\mathcal{R}_d^2 \\ \Leftrightarrow\quad & c(s,\eta)d^s\omega_n^2 b_n^d h_n^{2s-d}\|\psi_0\|_2^{2d} \leq \mathcal{R}_d^2. \\ \Leftrightarrow\quad & c(s,\eta)d^s\omega_n^2 b_n^{2s} h^{2s-d}\|\psi_0\|_2^{2d} \leq \mathcal{R}_d^2.\end{aligned} \tag{23}$$

L_2-separation Note that the L_2-distance between $p_0^{(n)}$ and $q_{a_n}^{(n)}$ can be computed as

$$\begin{aligned}
\|p_0^{(n)} - q_{a_n}^{(n)}\|_2^2 &= \left\| \omega_n \sum_{k \in \mathcal{K}_n} a_{n,k} \psi_{n,k}(x) \right\|_2^2 \\
&= \omega_n^2 \left\| \sum_{k \in \mathcal{K}_n} a_{n,k} \psi_{n,k}(x) \right\|_2^2 \\
&= \omega_n^2 \sum_{k \in \mathcal{K}_n} \|\psi_{n,k}(x)\|_2^2 \\
&= \omega_n^2 b_n^d \int \left|\psi_n(h_n x)\right|^2 dx \\
&= \omega_n^2 b_n^d h_n^{-d} \int |\psi_n(t)|^2 dt \quad (t := h_n x \in \mathbb{R}^d) \\
&= \omega_n^2 h^{-d} \|\psi_n\|_2^2,
\end{aligned}$$

where the second equality follows from the disjoint support of $\psi_{n,k}$ for different $k \in \mathcal{K}_n$. Thus,

$$\|p_0^{(n)} - q_{a_n}^{(n)}\|_2 = \omega_n h^{-d/2} \|\psi_0\|_2^d. \tag{24}$$

Resulting lower bound Now, we recall Equation (20), Equation (21) and Equation (23), and summarize the result with $M_d = M^d$:

$$\begin{aligned}
\text{(Risk control)} \qquad & b_n^{-d} h^{-4d} n^2 \omega_n^4 \|\psi_0\|_2^{4d} \le c(\alpha, \beta, \eta), \\
\text{(Validity)} \qquad & \omega_n \le e^d \min\{h^d, M^d - h^d\}, \\
\text{(Inclusion in Sobolev ball)} \qquad & c(s, \eta) d^s \omega_n^2 b_n^{2s} h^{2s-d} \|\psi_0\|_2^{2d} \le \mathcal{R}_d^2, \\
L_2\text{-separation} \qquad & \|p_0^{(n)} - q_{a_n}^{(n)}\|_2 = \omega_n h^{-d/2} \|\psi_0\|_2^d.
\end{aligned}$$

We optimize the choice of ω_n, b_n and h to maximize the resulting lower bound on the separation, i.e., the slowest possible decaying rate of $\|p_0^{(n)} - q_{a_n}^{(n)}\|_2$ that satisfies the above three conditions.

$$\begin{aligned}
\|\psi_0\|_2^2 &= \frac{1}{2} \int_{-1}^{1} \exp\left(-\frac{2}{1 - x^2} \right) dx \\
&\ge \frac{1}{2} \int_{|x| \le 1/2} \exp\left(-\frac{2}{1 - x^2} \right) dx.
\end{aligned}$$

For $|x| \le 1/2$, we have $1 - x^2 \ge 3/4$, and hence

$$\frac{2}{1 - x^2} \le \frac{8}{3}.$$

Therefore,

$$\exp\left(-\frac{2}{1 - x^2} \right) \ge e^{-8/3}.$$

Thus,

$$\begin{aligned}
\|\psi_0\|_2^2 &\ge \frac{1}{2} \int_{|x| \le 1/2} e^{-8/3} \, dx \\
&= \frac{1}{2} e^{-8/3}.
\end{aligned}$$

Consequently,

$$\|\psi_0\|_2^2 \geq \frac{1}{2}e^{-8/3}, \qquad \|\psi_0\|_2 \geq \frac{1}{\sqrt{2}}e^{-4/3}.$$

which implies that $e\|\psi_0\|_2 \geq 2^{-1/2}e^{-1/3} \approx 0.5066..$ holds. Let us denote $a = e\|\psi_0\|_2$, and $\epsilon = \|p_0^{(n)} - q_{a_n}^{(n)}\|_2 = \omega_n h^{-d/2}\|\psi_0\|_2^d$. Then, the above conditions can be rewritten as

$$\begin{aligned}
\text{(Risk control)} \qquad & b_n^{-d}n^2\epsilon^4 h^{-2d} \leq c(\alpha,\beta,\eta),\\
\text{(Validity)} \qquad & \epsilon \leq a^d h^{-d/2}\min\{h^d, M^d - h^d\},\\
\text{(Inclusion in Sobolev ball)} \quad & c(s,\eta)d^s\epsilon^2 b_n^{2s}h^{2s} \leq \mathcal{R}_d^2.
\end{aligned}$$

Choose $h = M2^{-1/d}$, then, the above conditions can be rewritten as

$$\begin{aligned}
\text{(Risk control)} \quad & \epsilon \leq c'(\alpha,\beta,\eta)b_n^{d/4}n^{-1/2}M^{d/2},\\
\text{(Validity)} \quad & \epsilon \leq ca^dM^{d/2},\\
\text{(Inclusion in Sobolev ball)} \quad & \epsilon \leq c'(s,\eta,M)d^{-s/2}b_n^{-s}M^{-s}2^{s/d}\mathcal{R}_d.
\end{aligned}$$

And therefore, this guarantees the minimax separation rate is lower bounded by

$$\epsilon_n^{(n)^*}(\beta,\mathcal{C}_1^{(n)}) \geq \sup_{b_n}\left\{c'(\alpha,\beta,\eta)b_n^{d/4}n^{-1/2}M^{d/2} \wedge ca^dM^{d/2} \wedge c'(s,\eta,M)d^{-s/2}b_n^{-s}2^{s/d}\mathcal{R}_d\right\}.$$

We balance the first and the third term:

$$b_n^{d/4}n^{-1/2}M^{d/2} = c(\alpha,\beta,\eta,s,M)d^{-s/2}b_n^{-s}M^{-s}2^{s/d}\mathcal{R}_d,$$

which yields

$$\begin{aligned}
b_n &= c(\alpha,\beta,\eta,s,M)^{\frac{4}{4s+d}}n^{\frac{2}{4s+d}}d^{-\frac{2s}{4s+d}}M^{-\frac{4s+2d}{4s+d}}2^{\frac{4s}{d(4s+d)}}\mathcal{R}_d^{\frac{4}{4s+d}},\\
b_n^{d/4} &= c(\alpha,\beta,\eta,s,M)^{\frac{d}{4s+d}}n^{\frac{d}{2(4s+d)}}d^{-\frac{ds}{2(4s+d)}}M^{-\frac{d(2s+d)}{2(4s+d)}}2^{\frac{s}{(4s+d)}}\mathcal{R}_d^{\frac{d}{4s+d}}
\end{aligned}$$

and plugging the above choice of b_n into the first term, we have

$$\epsilon_n^{(n)^*}(\beta,\mathcal{C}_1^{(n)}) \geq c'(\alpha,\beta,\eta)c(\alpha,\beta,\eta,s,M)^{\frac{d}{4s+d}}n^{-\frac{2s}{4s+d}}d^{-\frac{ds}{2(4s+d)}}M^{\frac{sd}{4s+d}}2^{\frac{s}{4s+d}}\mathcal{R}_d^{\frac{d}{4s+d}} \wedge ca^dM^{d/2},$$

and since $c(\alpha,\beta,\eta,s,M)^{\frac{d}{4s+d}} \geq \min\{1, c(\alpha,\beta,\eta,s,M)\}$, and $2^{\frac{s}{4s+d}} \geq 1$, the above lower bound can be simplified as

$$\epsilon_n^{(n)^*}(\beta,\mathcal{C}_1^{(n)}) \geq c''(\alpha,\beta,\eta,s,M)n^{-\frac{2s}{4s+d}}d^{-\frac{ds}{2(4s+d)}}M^{\frac{sd}{4s+d}}\mathcal{R}_d^{\frac{d}{4s+d}} \wedge ca^dM^{d/2},$$

Since $M > 4$ and $a = e\|\psi_0\|_2 \geq 0.5$, we have $a^dM^{d/2} \geq 1$. Therefore, we have

$$\epsilon_n^{(n)^*}(\beta,\mathcal{C}_1^{(n)}) \geq c''(\alpha,\beta,\eta,s,M)n^{-\frac{2s}{4s+d}}d^{-\frac{ds}{2(4s+d)}}M^{\frac{sd}{4s+d}}\mathcal{R}_d^{\frac{d}{4s+d}} \wedge c,$$

Note that $\mathcal{R}_d = \mathcal{R}d^{s/2}$, and the resulting lower bound can be rewritten as

$$\epsilon_n^{(n)^*}(\beta,\mathcal{C}_1^{(n)}) \geq c''(\alpha,\beta,\eta,s,M)n^{-\frac{2s}{4s+d}}M^{\frac{sd}{4s+d}}\mathcal{R}^{\frac{d}{4s+d}} \wedge c,$$

and n can be replaced by N since the sample size is balanced.

C.6.2. Upper bound analysis

In this section, we refer to the analysis in Schrab et al. (2023) and unpack the dependency of d in the resulting upper bound. Note that $n \leq m \leq Cn$ for some constant $C \geq 1$ is assumed in their analysis. We start from Schrab et al. (2023, Page 66), where they state that a sufficient condition for the MMD test to have type II error at most β:

$$\mathbb{P}\left(\|\psi\|_2^2 \geq \|\psi - \psi * \varphi_\lambda\|_2^2 - \|\psi * \varphi_\lambda\|^2 + 2\sqrt{\frac{2}{\beta}\mathrm{var}\left(\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}; k_\lambda)\right)} + 2q_{1-\alpha,B}(k_\lambda)\right) \geq 1 - \frac{\beta}{2},$$

and the quantile $q_{1-\alpha,B}(k_\lambda)$ has a high probability upper bound as

$$\mathbb{P}\left(q_{1-\alpha,B}(k_\lambda) \leq \tilde{C}_2(M_d, d, \alpha, \beta)\frac{1}{N\sqrt{\lambda_1 \cdots \lambda_d}}\right) \geq 1 - \frac{\beta}{2},$$

where $\tilde{C}_2(M_d, d, \alpha, \beta) = c\sqrt{M_d \kappa_2}$, and $c > 0$ is a universal constant. Therefore, it suffices to have

$$\|\psi\|_2^2 \geq \|\psi - \psi * \varphi_\lambda\|_2^2 - \|\psi * \varphi_\lambda\|^2 + 2\sqrt{\frac{2}{\beta}\mathrm{var}\left(\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}; k_\lambda)\right)} + 2\tilde{C}_2(M_d, d, \alpha, \beta)\frac{1}{N\sqrt{\lambda_1 \cdots \lambda_d}}.$$

From Schrab et al. (2023, Page 67) we have a variance upper bound as

$$2\sqrt{\frac{2}{\beta}\mathrm{var}\left(\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}; k_\lambda)\right)} \leq \|\psi * \varphi_\lambda\|^2 + c(\beta)\frac{2C_1(M_d, d)}{\beta N} + \frac{2\sqrt{C_1(M_d, d)}}{\sqrt{\beta} N\sqrt{\lambda_1 \cdots \lambda_d}},$$

a sufficient condition for the above inequality to hold is

$$\|\psi\|_2^2 - \|\psi - \psi * \varphi_\lambda\|_2^2 \geq c(\beta)\frac{2C_1(M_d, d)}{\beta N} + \frac{2\sqrt{C_1(M_d, d)}}{\sqrt{\beta} N\sqrt{\lambda_1 \cdots \lambda_d}} + 2\tilde{C}_2(M_d, d, \alpha, \beta)\frac{1}{N\sqrt{\lambda_1 \cdots \lambda_d}}.$$

For the normalized Gaussian product kernel, we have

$$\kappa_2 = \prod_{i=1}^{d} \int_{\mathbb{R}} K_i(u)^2 du = (2\pi)^{-d/2}.$$

Since M_d is the uniform upper bound on the densities, the constants satisfy

$$C_1(M_d, d) = \max\{2c_0^\dagger M_d, 16c_0^\dagger M_d \kappa_2\} \leq 16 M_d c_0^\dagger (1 + \kappa_2) =: C' M_d (1 + \kappa_2),$$

where $C' > 0$ depends only on the sample-size balance constant and the universal constants. Also,

$$\sqrt{C_1(M_d, d)} \leq \sqrt{C'} M_d^{1/2}, \qquad \tilde{C}_2(M_d, d, \alpha, \beta) = C(\alpha, \beta) M_d^{1/2} \kappa_2^{1/2},$$

for some constant $C(\alpha, \beta) > 0$ depending only on α and β. Therefore, the above sufficient condition is implied by

$$\|\psi\|_2^2 - \|\psi - \psi * \varphi_\lambda\|_2^2 \geq C(\alpha, \beta)\frac{M_d(1 + \kappa_2)}{N} + C(\alpha, \beta)\frac{M_d^{1/2}\kappa^{1/2}}{N\sqrt{\lambda_1 \cdots \lambda_d}},$$

where $C(\alpha, \beta) > 0$ is a constant depending only on α and β. Now take the common bandwidth $\lambda_1 = \cdots = \lambda_d = \lambda_0$. Then $\sqrt{\lambda_1 \cdots \lambda_d} = \lambda_0^{d/2}$, and the condition becomes

$$\|\psi\|_2^2 - \|\psi - \psi * \varphi_\lambda\|_2^2 \geq C(\alpha, \beta)\frac{M_d(1 + \kappa_2)}{N} + C(\alpha, \beta)\frac{M_d^{1/2}\kappa^{1/2}}{N\lambda_0^{d/2}}.$$

From Schrab et al. (2023, Page 69), we have

$$\|\psi - \psi * \varphi_\lambda\|_2^2 \leq S^2\|\psi\|_2^2 + T_s^2 \mathcal{R}^2 \|\lambda\|_2^{2s},$$

where

$$S(t) := \sup_{\|\xi\|_2 \le t} |1 - g(\xi)|, \qquad T_s(t) := \sup_{\|\xi\|_2 > t} \frac{|g(\xi)|}{\|\xi\|_2^s},$$

where $g(\xi) := 1 - \prod_{i=1}^d \widehat{K}_i(\xi_i)$.

Therefore, the sufficient condition is implied by

$$\|\psi\|_2^2 \ge C(\alpha,\beta)\frac{M_d(1+\kappa_2)}{N(1-S(t)^2)} + C(\alpha,\beta)\frac{M_d^{1/2}\kappa_2^{1/2}}{N\lambda_0^{d/2}(1-S(t)^2)} + C(\alpha,\beta)\frac{T_s(t)^2\mathcal{R}_d^2\|\lambda\|_2^{2s}}{1-S(t)^2}.$$

Gaussian Kernel For the (normalized) Gaussian kernel $K_i(u) = e^{-u^2}/\sqrt{\pi}$ for all $i \in [d]$, the squared L^2 norm κ_2 of the product gaussian kernel is

$$\kappa_2 = \left(\frac{1}{\pi}\int_{\mathbb{R}} e^{-2u^2} du\right)^d = (2\pi)^{-d/2}.$$

Since the sample sizes are balanced, there exists a constant $c_2 > 0$ such that the constant $C_3(M,d) \le c_2 M$. Also, the Fourier transform $\widehat{K}_i$ of K_i is

$$\widehat{K}_i(\xi_i) = \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} e^{-\xi_i^2 - ix\xi_i} dx = e^{-\xi_i^2/4}.$$

Therefore, $g(\xi) := 1 - \prod_{i=1}^d e^{-\xi_i^2/4} = 1 - e^{-\|\xi\|^2/4}$. This means for any $t > 0$,

$$S(t) = \sup_{\|\xi\|_2 \le t} g(\xi) = 1 - e^{-t^2/4}, \qquad T_s = \sup_{\|\xi\|_2 > t} \frac{1 - e^{-\|\xi\|_2^2/4}}{\|\xi\|_2^s} \le \sup_{\|\xi\|_2 > t} \frac{1}{\|\xi\|_2^s} = t^{-s}.$$

Using the Gaussian expressions above, we have $S(t) = 1 - e^{-t^2/4}$ and $T_s(t) \le t^{-s}$. Moreover,

$$\frac{1}{1-S(t)^2} = \frac{1}{1-(1-e^{-t^2/4})^2} = \frac{e^{t^2/4}}{2 - e^{-t^2/4}} \le e^{t^2/4}.$$

Therefore, the previous sufficient condition is implied by

$$\|\psi\|_2^2 \ge C(\alpha,\beta)e^{t^2/4}\left(\frac{M_d(1+\kappa_2)}{N} + \frac{M_d^{1/2}\kappa_2^{1/2}}{N\lambda_0^{d/2}} + t^{-2s}\mathcal{R}_d^2\|\lambda\|_2^{2s}\right).$$

Since $\lambda_1 = \cdots = \lambda_d = \lambda_0$, we have $\|\lambda\|_2^{2s} = d^s\lambda_0^{2s}$. Also, for the normalized Gaussian product kernel, $\kappa_2 = (2\pi)^{-d/2} \le 1$. Hence, for any fixed $t > 0$,

$$\|\psi\|_2^2 \ge C(\alpha,\beta)e^{t^2/4}\left(\frac{M_d}{N} + \frac{M_d^{1/2}}{N\lambda_0^{d/2}} + t^{-2s}\mathcal{R}_d^2 d^s\lambda_0^{2s}\right)$$

is sufficient. Taking, for example, $t = 2\sqrt{s}$, the factor $e^{t^2/4}t^{-2s} = e^s(2\sqrt{s})^{-2s}$ depends only on s. Thus there exists a constant $C(\alpha,\beta,s) > 0$ such that it is sufficient to have

$$\|\psi\|_2^2 \ge C(\alpha,\beta,s)\left(\mathcal{R}_d^2 d^s\lambda_0^{2s} + \frac{M_d}{N} + \frac{M_d^{1/2}}{N\lambda_0^{d/2}}\right).$$

Consequently, the squared separation radius satisfies

$$\rho_n(\Delta_\alpha^{B,\lambda};\beta,\mathcal{C}_1^{(n)})^2 \le C(\alpha,\beta,s)\left(\mathcal{R}_d^2 d^s\lambda_0^{2s} + \frac{M_d}{N} + \frac{M_d^{1/2}}{N\lambda_0^{d/2}}\right).$$

We choose λ_0 by balancing the first and third terms:

$$\mathcal{R}_d^2 d^s \lambda_0^{2s} = \frac{M_d^{1/2}}{N\lambda_0^{d/2}}.$$

Equivalently,

$$\lambda_0^{2s+d/2} = \frac{M_d^{1/2}}{\mathcal{R}_d^2 d^s N}, \qquad \lambda_0^\star = \left(\frac{M_d^{1/2}}{\mathcal{R}_d^2 d^s N}\right)^{2/(4s+d)}.$$

At this bandwidth,

$$\mathcal{R}_d^2 d^s (\lambda_0^\star)^{2s} = \mathcal{R}_d^2 d^s \left(\frac{M_d^{1/2}}{\mathcal{R}_d^2 d^s N}\right)^{4s/(4s+d)} = \mathcal{R}_d^{2d/(4s+d)} d^{sd/(4s+d)} M_d^{2s/(4s+d)} N^{-4s/(4s+d)}.$$

Since the first and third terms are balanced, substituting $\lambda_0^\star$ gives

$$\rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)})^2 \le C(\alpha,\beta,s)\left(\mathcal{R}_d^{2d/(4s+d)} d^{sd/(4s+d)} M_d^{2s/(4s+d)} N^{-4s/(4s+d)} + \frac{M_d}{N}\right)$$

which yields

$$\rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)}) \le C(\alpha,\beta,s)\left(\mathcal{R}_d^{d/(4s+d)} d^{sd/(2(4s+d))} M_d^{s/(4s+d)} N^{-2s/(4s+d)} + M_d^{1/2} N^{-1/2}\right),$$

where we use $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for any $x, y \ge 0$. Note that $\mathcal{R}_d = \mathcal{R} d^{s/2}$ and $M_d = M^d$, and it is equivalent to

$$\rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)}) \le C(\alpha,\beta,s)\left(\mathcal{R}^{d/(4s+d)} M^{sd/(4s+d)} N^{-2s/(4s+d)} + M^{d/2} N^{-1/2}\right).$$

Laplace Kernel For the normalized Laplace kernel $K_i(u) = e^{-|u|}/2$ for all $i \in [d]$, the Fourier transform is

$$\widehat{K}_i(\xi_i) = \frac{1}{2}\int_{\mathbb{R}} e^{-|u|} e^{-iu\xi_i} du = \frac{1}{1+\xi_i^2}.$$

Therefore,

$$g(\xi) = 1 - \prod_{i=1}^d \frac{1}{1+\xi_i^2}.$$

For any $t > 0$, since $\sum_{i=1}^d \xi_i^2 \le t^2$ and $a \mapsto \log(1+a)$ is concave on $[0,\infty)$, Jensen's inequality gives

$$\prod_{i=1}^d (1+\xi_i^2) \le \left(1 + \frac{t^2}{d}\right)^d.$$

Hence,

$$S(t) = \sup_{\|\xi\|_2 \le t}\left(1 - \prod_{i=1}^d \frac{1}{1+\xi_i^2}\right) = 1 - \left(1 + \frac{t^2}{d}\right)^{-d}.$$

Also, since $0 \le g(\xi) \le 1$, we have

$$T_s(t) = \sup_{\|\xi\|_2 > t} \frac{|g(\xi)|}{\|\xi\|_2^s} \le t^{-s}.$$

Now,

$$\frac{1}{1-S(t)^2} = \frac{1}{1-\left(1-\left(1+\frac{t^2}{d}\right)^{-d}\right)^2} = \frac{\left(1+\frac{t^2}{d}\right)^d}{2-\left(1+\frac{t^2}{d}\right)^{-d}} \le \left(1+\frac{t^2}{d}\right)^d \le e^{t^2}.$$

Moreover, the squared L^2 norm of the one-dimensional normalized Laplace kernel is

$$\int_{\mathbb{R}} K_i(u)^2 du = \frac{1}{4}\int_{\mathbb{R}} e^{-2|u|} du = \frac{1}{4},$$

and hence

$$\kappa_2 = \prod_{i=1}^{d} \int_{\mathbb{R}} K_i(u)^2 du = 4^{-d}.$$

Therefore, the previous sufficient condition is implied by

$$\|\psi\|_2^2 \ge C(\alpha,\beta)e^{t^2}\left(\frac{M_d(1+\kappa_2)}{N} + \frac{M_d^{1/2}\kappa_2^{1/2}}{N\lambda_0^{d/2}} + t^{-2s}\mathcal{R}_d^2\|\lambda\|_2^{2s}\right).$$

Since $\lambda_1 = \cdots = \lambda_d = \lambda_0$, we have $\|\lambda\|_2^{2s} = d^s\lambda_0^{2s}$. Also, since $\kappa_2 = 4^{-d} \le 1$, for any fixed $t > 0$,

$$\|\psi\|_2^2 \ge C(\alpha,\beta)e^{t^2}\left(\frac{M_d}{N} + \frac{M_d^{1/2}}{N\lambda_0^{d/2}} + t^{-2s}\mathcal{R}_d^2 d^s\lambda_0^{2s}\right)$$

is sufficient. Taking, for example, $t = \sqrt{s}$, the factor $e^{t^2}t^{-2s} = e^s s^{-s}$ depends only on s. Thus there exists a constant $C(\alpha,\beta,s) > 0$ such that it is sufficient to have

$$\|\psi\|_2^2 \ge C(\alpha,\beta,s)\left(\mathcal{R}_d^2 d^s\lambda_0^{2s} + \frac{M_d}{N} + \frac{M_d^{1/2}}{N\lambda_0^{d/2}}\right).$$

Consequently, the squared separation radius satisfies

$$\rho_n(\Delta_\alpha^{B,\lambda};\beta,\mathcal{C}_1^{(n)})^2 \le C(\alpha,\beta,s)\left(\mathcal{R}_d^2 d^s\lambda_0^{2s} + \frac{M_d}{N} + \frac{M_d^{1/2}}{N\lambda_0^{d/2}}\right).$$

We choose λ_0 by balancing the first and third terms:

$$\mathcal{R}_d^2 d^s\lambda_0^{2s} = \frac{M_d^{1/2}}{N\lambda_0^{d/2}}.$$

Equivalently,

$$\lambda_0^{2s+d/2} = \frac{M_d^{1/2}}{\mathcal{R}_d^2 d^s N}, \qquad \lambda_0^\star = \left(\frac{M_d^{1/2}}{\mathcal{R}_d^2 d^s N}\right)^{2/(4s+d)}.$$

At this bandwidth,

$$\mathcal{R}_d^2 d^s(\lambda_0^\star)^{2s} = \mathcal{R}_d^2 d^s\left(\frac{M_d^{1/2}}{\mathcal{R}_d^2 d^s N}\right)^{4s/(4s+d)} = \mathcal{R}_d^{2d/(4s+d)} d^{sd/(4s+d)} M_d^{2s/(4s+d)} N^{-4s/(4s+d)}.$$

Since the first and third terms are balanced, substituting $\lambda_0^\star$ gives

$$\rho_n(\Delta_\alpha^{B,\lambda^*};\beta,\mathcal{C}_1^{(n)})^2 \le C(\alpha,\beta,s)\left(\mathcal{R}_d^{2d/(4s+d)} d^{sd/(4s+d)} M_d^{2s/(4s+d)} N^{-4s/(4s+d)} + \frac{M_d}{N}\right).$$

Taking square roots and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, we obtain

$$\rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)}) \le C(\alpha, \beta, s) \left(\mathcal{R}_d^{d/(4s+d)} d^{sd/(2(4s+d))} M_d^{s/(4s+d)} N^{-2s/(4s+d)} + M_d^{1/2} N^{-1/2} \right).$$

Note that $\mathcal{R}_d = \mathcal{R} d^{s/2}$ and $M_d = M^d$, and it is equivalent to

$$\rho_n(\Delta_\alpha^{B,\lambda^*}; \beta, \mathcal{C}_1^{(n)}) \le C(\alpha, \beta, s) \left(\mathcal{R}^{d/(4s+d)} M^{sd/(4s+d)} N^{-2s/(4s+d)} + M^{d/2} N^{-1/2} \right),$$

D. Time and memory savings in the Fenwick tree algorithm

D.1. Fast algorithm with Fenwick structure

We take the one-dimensional case as a pedagogical example to illustrate the main idea of the algorithm. Suppose that we are computing the lower triangular sum $\sum_{j<i} k(z_i, z_j)$, where a kernel k is PS. First, as a preprocessing step, we identify the order of the data points $z_1, z_2, \ldots, z_N$ within $O(N \log N)$ time, and denote the rank of z_i as r_i. At i-th step, we split the row sum into two parts:

$$C_i = \sum_{j<i} k(z_i, z_j) = f_-(z_i) \sum_{j<i} \mathbf{1}\{z_j \le z_i\} g_-(z_j) + f_+(z_i) \sum_{j<i} \mathbf{1}\{z_j > z_i\} g_+(z_j).$$

Assume that there exists a data structure `T=(T[1],...,T[N])` which has two $O(\log n)$ time operations: (i) $\texttt{T.update}(a, v)$ which increments the a-th element of `T` by v, i.e., `T[a]` $\leftarrow$ `T[a]` $+ v$; (ii) $\texttt{T.query}(a)$ which returns the a-th prefix sum $\sum_{k=1}^{a} \texttt{T[k]}$. Now, we define two data structures $\texttt{T}_-$ and $\texttt{T}_+$ to store the prefix sums of $g_-(z_j)$ and $g_+(z_j)$, respectively. Initially, we set $\texttt{T}_- = \texttt{T}_+ = \texttt{(0,...,0)}$. At i-th step, we first run queries and get the outputs $G_{i,-} = \texttt{T}_-\texttt{.query}(r_i)$, $G_{i,+} = \texttt{T}_+\texttt{.query}(r_i)$, and $G_{N,+} = \texttt{T}_+\texttt{.query}(N)$. Then compute $C_i = f_-(z_i) G_{i,-} + f_+(z_i)(G_{N,+} - G_{i,+})$. Note that $G_{N,+} - G_{i,+}$ gives suffix sum $\sum_{k=r_i+1}^{N} \texttt{T}_+\texttt{[k]}$. Finally, we update the data structures by $\texttt{T}_-\texttt{.update}(r_i, g_-(z_i))$ and $\texttt{T}_+\texttt{.update}(r_i, g_+(z_i))$. Correctness of the algorithm follows by verifying that $G_{i,-} := \sum_{k=1}^{r_i} \texttt{T}_-\texttt{[k]}$ and $\sum_{j<i} \mathbf{1}\{z_j \le z_i\} g_-(z_j)$ are the same, after $i-1$ updates. Indeed, for any $t = 1, \ldots, i-1$, note that r_t-th element of $\texttt{T}_-$ updated to $g_-(z_t)$ after t-th update, and the other elements remain zero. Then at i-th query step, we have $G_{i,-} = \sum_{k=1}^{r_i} \texttt{T}_-\texttt{[k]}$, provided that for each $k \le r_i$, `[k]` is either zero or $g_-(z_t)$ for some $t \le i-1$ with $r_t = k$. A similar argument applies to $G_{i,+}$ and $G_{N,+}$ as well, and one can verify that C_i is correctly computed.

A generalization of the above algorithm is straightforward for the multi-dimensional case. Suppose that there exist a multidimensional array structure $\texttt{T} = (\texttt{T[k]})_{k \in \{1,\ldots,N\}^d}$ which can be updated and queried as follows: for $a = (a_1, \ldots, a_d)$, we define query for a-th box prefix sum as $\texttt{T.query}(a) := \sum_{k_1=1}^{a_1} \cdots \sum_{k_d=1}^{a_d} \texttt{T}[k_1, \ldots, k_d]$, and the update operation is defined as $\texttt{T.update}(a, v)$ which updates the a-th element of `T` as `T[a]` $+ v$. Note that at step i, the algorithm satisfies $\texttt{T.query}(r_i) = \sum_{j<i} \mathbf{1}\{r_{j,1} \le r_{i,1}, \ldots, r_{j,d} \le r_{i,d}\} g(z_j)$, where $r_{i,k}$, $r_{j,k}$ are the coordinate-wise rank of z_i, z_j in the k-th dimension, respectively. Now, consider a lower kernel triangular sum $K_{ZZ} = \sum_{1 \le j < i \le N} k(z_i, z_j)$ where k is OS. For each $2^d = |\{s : s \in \{+,-\}^d\}|$ orthants, we define $\texttt{T}_s$ to store the d-dimensional box prefix sums of $g_s(z_j)$. Then at i-th step, we run queries on all $\texttt{T}_s$ to get the prefix sums, and compute

$$C_i = \sum_{s \in \{+,-\}^d} f_s(z_i) \sum_{j<i} \mathbf{1}\{z_j \in Q_s(z_i)\} g_s(z_j),$$

by inclusion-exclusion with box prefix sums, and update $\texttt{T}_s$.

D.2. Algorithm pseudocode and Fenwick structure

We provide a naive implementation of kernel lower triangular sum, given the proper structure supporting query and update operations.

Algorithm 1 $\mathrm{FastKernelSum}(\mathbf{z}_N, \mathbf{r}_N, k)$; $\odot$ denotes element-wise product

Require: Samples $\mathbf{z}_N = (z_1, \ldots, z_N)$, coordinate-wise ranks $\mathbf{r}_N = (r_1, \ldots, r_N)$, OS kernel k
Ensure: $K = \sum_{1 \le j < i \le N} k(z_i, z_j)$
  $K \leftarrow 0$
  for $s \in \{+, -\}^d$ **do**
    Initialize a d-dimensional array $\mathtt{T}_s$ with all entries zero
    Let $\mathbf{1}_s^+ = (\mathbf{1}\{s_1 = +\}, \ldots, \mathbf{1}\{s_d = +\})$ and $\mathbf{1}_s^- = \mathbf{1} - \mathbf{1}_s^+$
  end for
  for $i = 1, \ldots, N$ **do**
    $C_i \leftarrow 0$
    for $s \in \{+, -\}^d$ **do**
      $a_i^{(s)} \leftarrow \mathbf{1}_s^+ \odot (r_i + \mathbf{1}) + \mathbf{1}_s^-, \qquad b_i^{(s)} \leftarrow \mathbf{1}_s^- \odot r_i + \mathbf{1}_s^+ \odot N\mathbf{1}$
      $G_{i,s} \leftarrow \sum_{\epsilon \in \{0,1\}^d} (-1)^{|\epsilon|} \mathtt{T}_s.\mathtt{query}\Big(\epsilon \odot (a_i^{(s)} - \mathbf{1}) + (1 - \epsilon) \odot b_i^{(s)}\Big)$
      $C_i \leftarrow C_i + f_s(z_i) G_{i,s}$
    end for
    $K \leftarrow K + C_i$
    for $s \in \{+, -\}^d$ **do**
      $\mathtt{T}_s.\mathtt{update}(r_i, g_s(z_i))$
    end for
  end for
  return K

Fenwick structure. One of the data structures that satisfies the above requirements is the Fenwick structure (Fenwick, 1994), which supports update and query operations in $O(\log N)$ time for both update and query operations when $d = 1$. At the heart of the Fenwick structure is to store the data as a proper partition of the prefix sums. Instead of storing the original data $\mathtt{T[1]}, \ldots, \mathtt{T[N]}$, the Fenwick structure stores the data as partial sums $\mathtt{p} = (\mathtt{p[1]}, \ldots, \mathtt{p[N]})$ where $\mathtt{p[i]} = \sum_{k=i-2^r+1}^{i} \mathtt{T[k]}$ where r is the largest integer such that 2^r divides i. This enables faster prefix sum, for example, $\sum_{k=1}^{7} \mathtt{T[k]}$ can be computed by $\mathtt{p[4]} + \mathtt{p[6]} + \mathtt{p[7]}$, where $\mathtt{p[4]} = \sum_{k=1}^{4} \mathtt{T[k]}$, $\mathtt{p[6]} = \sum_{k=5}^{6} \mathtt{T[k]}$ and $\mathtt{p[7]} = \mathtt{T[7]}$. The update operation can be done by updating the corresponding partial sums, for example, if we update $\mathtt{T[3]}$ by v, we need to update all partial sums that include $\mathtt{T[3]}$, which are $\mathtt{p[3]}, \mathtt{p[4]}, \mathtt{p[8]}, \ldots$. A similar idea can be applied to the multi-dimensional case with time complexity $O(\log^d N)$ for both update and query operations as well (Mishra, 2013; Wang et al., 2021), where the data is now stored as box shape partial sums. To be more specific, the d-dimensional Fenwick structure stores partial sums $\mathtt{p}[\mathbf{a}]$ corresponding to axis-aligned boxes that are Cartesian products of 1D Fenwick structures:

$$\mathtt{p}[a_1, \ldots, a_d] = \sum_{k_1 = a_1 - 2^{r_1} + 1}^{a_1} \cdots \sum_{k_d = a_d - 2^{r_d} + 1}^{a_d} \mathtt{T}[k_1, \ldots, k_d],$$

where, for each coordinate $l \in [d]$, r_l is the largest integer such that 2^{r_l} divides a_l. A tree-based implementation is presented in Appendix D.4.

D.3. Improved time complexity

A naive implementation of Algorithm 1 has time complexity $O(4^d N \log^d N)$ as follows: we first identify the coordinate-wise rank of each data point in $O(dN \log N)$ time. Then for each 2^d orthants, we need at most 2^d queries to compute C_i at i-th step, and each query has time complexity $O(\log^d N)$. In Appendix D.3, we present two techniques to further accelerate the algorithm. First, we claim that for symmetric kernels, the effective size of dimensions becomes $d - 1$ by feeding the data in the algorithm by their ranks in the first dimension, which allows us to save one logarithmic exponent. Second, we avoid inclusion-exclusion by considering auxiliary ranks which align with the orthant structure, so that we can compute the box sum by a single query. These two improvements reduces the overall time complexity of the algorithm as $O(2^d N \log^{(d-1)\vee 1} N)$. Also, one may be concerned that the Fenwick structure in Algorithm 1 as multidimensional array requires $O(N^d)$ memory, which is prohibitive for large N and d. In fact, the tree-based implementation can be done with

$O(2^{d-1}N\log^{(d-2)\vee 0}N)$ memory, based on the observation that most of the entries in the N^d-size tensors are zero and we only need to store the non-zero entries. We provide details below.

Saving one dimension in the Fenwick tree In this section, we claim that 2^{d-1} number of $(d-1)$-Fenwick trees are enough for algorithm. A key observation is that for symmetric kernels, the kernel lower triangular sum

$$K_{ZZ} = \sum_{j<i} k(z_i, z_j)$$

does not change even if we permute the order of the samples. Let π be a permutation such that $r_{\pi(1),1} \le r_{\pi(2),1} \le \cdots \le r_{\pi(n),1}$. Then, we feed the samples in the order of $z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(n)}$ to the Fenwick tree algorithm, and this does not change the value of K_{ZZ}. Therefore, in this section, we assume that $z_1, z_2, \ldots, z_n$ are ordered as $r_{1,1} \le r_{2,1} \le \cdots \le r_{n,1}$ without loss of generality.

Second, we note that for any $u < t$, we have $\mathrm{sign}(z_{u,1} - z_{t,1}) = -$, which implies that the orthant indicator $\mathbf{1}\{z_j \in Q_s(z_i)\}$ is zero for any s with $s_1 = +$. Therefore, we only need to consider the orthants with $s_1 = -$, which reduces the number of Fenwick trees from 2^d to 2^{d-1}, and also the dimensionality of the Fenwick tree from d to $d-1$. Indeed, define the sign vector as

$$(-, s') \in \{+,-\}^d, \quad \text{where } s' \in \{+,-\}^{d-1},$$

then the row sum becomes

$$C_i = \sum_{s' \in \{+,-\}^{d-1}} f_{(-,s')}(z_i) \sum_{j<i} \mathbf{1}\{\mathrm{sign}(z_{j,l} - z_{i,l}) = s'_l,\ l = 2, \ldots, d\} g_{(-,s')}(z_j).$$

Observe that we only need to maintain 2^{d-1} Fenwick trees $\mathtt{T}_{(-,s')}$ for $s' \in \{+,-\}^{d-1}$. We query as $G_{i,(-,s')} = \mathtt{T}_{(-,s')}\mathtt{.query}(r')$, and update as $\mathtt{T}_{(-,s')}\mathtt{.update}(r', g_{(-,s')}(z_i))$, where $r' = (r_{i,2}, \ldots, r_{i,d})$ is the rank vector of z_i excluding the first dimension.

Avoiding inclusion-exclusion principle in computing C_i Note that some of summands in C_i are not mere prefix sums, but also include suffix sums, which requires an extra $O(2^d)$ time complexity to run inclusion-exclusion principle. For example, consider the orthant $s = \{+,+,+\}$. We need to compute

$$f_s(z_i) \sum_{j<i} \mathbf{1}\{r_{j,1} > r_{i,1},\ r_{j,2} > r_{i,2},\ r_{j,3} > r_{i,3}\} g_s(z_j).$$

Define the prefix sum

$$F(a_1, a_2, a_3) := \sum_{j<i} \mathbf{1}\{r_{j,1} \le a_1,\ r_{j,2} \le a_2,\ r_{j,3} \le a_3\}\, g_s(z_j)$$

By the inclusion–exclusion principle, we calculate the suffix sum as

$$\begin{aligned}
&\sum_{j<i} \mathbf{1}\{r_{j,1} > r_{i,1},\ r_{j,2} > r_{i,2},\ r_{j,3} > r_{i,3}\} g_s(z_j) \\
&= F(N,N,N) - F(r_{i,1},N,N) - F(N,r_{i,2},N) - F(N,N,r_{i,3}) \\
&\quad + F(r_{i,1},r_{i,2},N) + F(r_{i,1},N,r_{i,3}) + F(N,r_{i,2},r_{i,3}) - F(r_{i,1},r_{i,2},r_{i,3}),
\end{aligned}$$

which takes 2^3 time. To avoid such $O(2^d)$ extra time complexity, we introduce an auxiliary rank $\bar{r}^{(s)}_{j,l}$ and query boundary $b^{(s)}_{i,l}$ which align to each orthant's ordering, so that the suffix sum can be computed by a single query:

$$\bar{r}^{(s)}_{j,l} = \begin{cases} r_{j,l}, & s_l = -, \\ N - r_{j,l} + 1, & s_l = +, \end{cases} \quad b^{(s)}_{i,l} = \begin{cases} r_{i,l}, & s_l = -, \\ N - r_{i,l}, & s_l = +. \end{cases}$$

Then, for each i-th step, we query as $G_{i,s} = \mathtt{T}_s\mathtt{.query}(b^{(s)}_{i,1}, \ldots, b^{(s)}_{i,d})$, calculate $C_i = \sum_{s\in\{+,-\}^d} f_s(z_i) G_{i,s}$, and update as $\mathtt{T}_s\mathtt{.update}(\bar{r}^{(s)}_i, g_s(z_i))$. We note that after $i-1$-updates, $\mathtt{T}_s\mathtt{.query}(b^{(s)}_{i,1}, \ldots, b^{(s)}_{i,d})$ gives the correct summation

$\sum_{j<i} \mathbf{1}\{z_j \in Q_s(z_i)\} g_s(z_j)$ for each $s \in \{+,-\}^d$. In summary, per each i-th step and each $s \in \{+,-\}^d$, we preprocess auxiliary ranks and query boundaries within $O(d)$ time and perform a single query in $O(\log^d N)$ time. This yields $O(d + \log^d N)$ time complexity per each (i, s)-pair, which is much more efficient than $O(2^d \log^d n)$ time complexity by the inclusion-exclusion principle.

We summarize the above two improvements in Algorithm 2, which is a more efficient version of Algorithm 1.

Algorithm 2 FastKernelSum$(\mathbf{z}_N, \mathbf{r}_N, k, d)$

Require: $\mathbf{z}_N = (z_1, \dots, z_N)$; axis-wise ranks $\mathbf{r}_N = (r_1, \dots, r_N)$; OS kernel k; dimension d.
Ensure: $K = \sum_{1 \le j < i \le N} k(z_i, z_j)$

```
Sort and relabel the samples so that r_{1,1} ≤ ··· ≤ r_{N,1}.
K ← 0
if d = 1 then
    G ← 0
    for i = 1, ..., N do
        K ← K + f_−(z_i) G
        G ← G + g_−(z_i)
    end for
    return K
end if
for s' ∈ {+, −}^{d−1} do
    s ← (−, s')
    Define auxiliary ranks and boundaries, for l = 2, ..., d,
```

$$\rho_{i,l}^{(s)} = \begin{cases} r_{i,l}, & s_l = -, \\ N - r_{i,l} + 1, & s_l = +, \end{cases} \qquad b_{i,l}^{(s)} = \begin{cases} \rho_{i,l}^{(s)}, & s_l = -, \\ \rho_{i,l}^{(s)} - 1, & s_l = +. \end{cases}$$

```
    Let ρ_i^(s) = (ρ_{i,2}^(s), ..., ρ_{i,d}^(s)).
    Let b_i^(s) = (b_{i,2}^(s), ..., b_{i,d}^(s)).
    Initialize a (d − 1)-dimensional Fenwick array T with all entries zero.
    for i = 1, ..., N do
        G_{i,s} ← T.query(b_i^(s))
        K ← K + f_s(z_i) G_{i,s}
        T.update(ρ_i^(s), g_s(z_i))
    end for
end for
return K
```

D.4. Memory-efficient implementation of the multi-dimensional Fenwick tree

An efficient way to implement the multi-dimensional Fenwick structure is to use a compressed Fenwick tree, which is designed for the coordinates that are actually updated. In other words, after identifying the coordinates that will be updated, we can build a nested tree structure that is specialized for such fixed rank geometry, preserving the same query and update time complexity but with much less memory usage. To display the structure more clearly, we give $d = 3$ case as an example instead of formal d-dimensional case. We will discuss the correctness of this compressed Fenwick tree in the end of this section. Also, we provide formal algorithms for building the compressed Fenwick tree, and performing query and update operations in this compressed Fenwick tree in Algorithms 5, 6, and 7 respectively.

Note that by introducing auxiliary ranks as in Appendix D.3, Computing C_i boils down to a single prefix sum query, so we restrict our attention to the prefix sum query in this example. First, let us define

$$\texttt{lowbit}(t) := 2^{\nu_2(t)},$$

where $\nu_2(t)$ is the largest integer such that $2^{\nu_2(t)}$ divides t. For example, $\texttt{lowbit}(12) = 4$ since $12 = 2^2 \cdot 3$ and $\nu_2(12) = 2$.

Also, define $\mathtt{I}(t) := \{t - \mathtt{lowbit}(t) + 1, \ldots, t\}$, which is the Fenwick block of t. For example, $\mathtt{I}(12) = \{9, 10, 11, 12\}$. We briefly introduce the idea of tree-shape implementation first. For d-dimensional prefix sum, a naive Fenwick algorithm saves the data as the summation of datapoints in the box $\mathtt{I}(a_1) \times \mathtt{I}(a_2) \times \cdots \times \mathtt{I}(a_d)$, which corresponds to

$$\mathtt{p}[a_1, \ldots, a_d] = \sum_{k_1 \in \mathtt{I}(a_1)} \cdots \sum_{k_d \in \mathtt{I}(a_d)} \mathtt{T}[k_1, \ldots, k_d],$$

for all $a_1, \ldots, a_d \in \{1, \ldots, N\}$. To calculate the prefix sum at $(b_1, \ldots, b_d)$, Fenwick algorithm returns it by finding a proper $A_1, \ldots, A_d$ such that

$$\sum_{k_1 \le b_1} \cdots \sum_{k_d \le b_d} \mathtt{T}[k_1, \ldots, k_d] = \sum_{a_1 \in A_1} \cdots \sum_{a_d \in A_d} \mathtt{p}[a_1, \ldots, a_d].$$

A memory bottleneck of this approach is the size of the multi-dimensional array $\mathtt{T}$, which is $O(N^d)$ in the worst case. In practice, only N of N^d coordinates are nonempty. Thus, a nested structure algorithm takes $(a_1, \ldots, a_d)$ and N nonempty coordinates as input and returns the exact sum $\mathtt{p}[a_1, \ldots, a_d]$ by performing iterative queries across dimensions 1 through d. For example, for each $a_1 \in A_1$, we ask which instances among N nonempty coordinates have their first coordinate in $\mathtt{I}(a_1)$. With the answer to this question, calculating $\sum_{k_1 \in \mathtt{I}(a_1)} \cdots \sum_{k_d \in \mathtt{I}(a_d)} \mathtt{T}[k_1, \ldots, k_d]$ boils down to sum of datapoints satisfying $k_1 \in \mathtt{I}(a_1)$ in the box $\mathtt{I}(a_2) \times \cdots \times \mathtt{I}(a_d)$, which is $(d-1)$-dimensional problem. Applying this dimension-query exchange idea recursively is the key idea of the nested structure algorithm, which alleviates the memory dependency on N^d.

Building Trees. Now, we shall think $\mathtt{T}$ as a nested tree structre, other than a multi-dimensional array $\mathtt{T} = \{\mathtt{T}[k_1, \ldots, k_d]\}$. Let us write $\mathtt{T}$ for a compressed three-dimensional Fenwick tree, which we will explain its construction in detail below. Suppose that $\mathtt{T.geometry}$ be the set of possible update locations in the compressed tree:

$$\mathtt{T.geometry} = \{(2, 8, 5),\ (4, 1, 7),\ (7, 6, 2),\ (9, 3, 9),\ (12, 5, 4)\},$$

and the values inserted at these locations be $v_1,\ v_2,\ v_3,\ v_4,\ v_5$, respectively. Let us define a function $\mathtt{T.coords}$ which returns the compressed coordinate list, which is the sorted list of first coordinates of elements in $\mathtt{T.geometry}$:

$$\mathtt{T.coords} = (2, 4, 7, 9, 12),$$

Note that the local compressed index, or equivalently the relative rank (RR) of the first coordinates among the compressed coordinates is labelled as

$$RR(2) = 1, \qquad RR(4) = 2, \qquad RR(7) = 3, \qquad RR(9) = 4, \qquad RR(12) = 5.$$

For a Fenwick node indexed by local compressed index q, define its one-dimensional Fenwick block by $\mathtt{I}(q) = \{q - \mathtt{lowbit}(q) + 1, \ldots, q\}$. Each top-level Fenwick node q stores a compressed two-dimensional Fenwick tree over the remaining coordinates (y, z) of all update locations whose compressed first-coordinate index lies in $\mathtt{I}(q)$. To be more precise,

q	$\mathtt{I}(q)$	points stored in $\mathtt{T.child}[q]$ as (y, z)
1	$\{1\}$	$\{(8, 5)\}$
2	$\{1, 2\}$	$\{(8, 5), (1, 7)\}$
3	$\{3\}$	$\{(6, 2)\}$
4	$\{1, 2, 3, 4\}$	$\{(8, 5), (1, 7), (6, 2), (3, 9)\}$
5	$\{5\}$	$\{(5, 4)\}$.

For example, $\mathtt{T.child}[4]$ stores all update locations whose local compressed index belongs to $\mathtt{I}(4) = \{1, 2, 3, 4\}$. This corresponds to first-coordinate values $2,\ 4,\ 7,\ 9$, and hence $\mathtt{T.child}[4]$ is a compressed two-dimensional Fenwick tree built from

$$\mathtt{T.child}[4]\mathtt{.geometry} = \{(8, 5), (1, 7), (6, 2), (3, 9)\}.$$

Now consider the internal structure of $\mathtt{T.child}[4]$. The sorted unique secondcoordinates are

$$\mathtt{T.child}[4]\mathtt{.coords} = (1, 3, 6, 8).$$

Thus, inside `T.child`[4], the local compressed indices are labelled as

$$RR(1) = 1, \qquad RR(3) = 2, \qquad RR(6) = 3, \qquad RR(8) = 4.$$

Each such node stores a one-dimensional compressed Fenwick tree over the remaining coordinate z. Therefore, inside `T.child`[4], the second-level structure is

q	$\mathtt{I}(q)$	points stored in `T.child`[4]`.child`[q] as (z)
1	$\{1\}$	$\{7\}$
2	$\{1, 2\}$	$\{7, 9\}$
3	$\{3\}$	$\{2\}$
4	$\{1, 2, 3, 4\}$	$\{2, 5, 7, 9\}$.

Note that a child of `T.child`[4] for each q is a one-dimensional compressed Fenwick tree, which is a simple array of values indexed by the compressed z-coordinates. For example, inside `T.child`[4]`.child`[2], the compressed z-coordinates are $\{7, 9\}$, and the corresponding compressed indices are $1, 2$, which are stored as `T.child`[4]`.child`[2]`.coords` $= (7, 9)$. Note that `T.child`[4]`.child`[2] is a one-dimensional Fenwick tree of length 2, so it stores the cumulative sums of the values inserted at the corresponding coordinates as one dimensional case. We stop branching when we reach the last dimension, or when a child tree has only one point, in which case we can directly store the value at the corresponding coordinate without building a Fenwick tree. For example, `T.child`[4]`.child`[3] stores only one point $(6, 2)$, so we directly store the value inserted at this point without building a one-dimensional Fenwick tree. We provide a formal algorithm for building the compressed Fenwick tree in Algorithm 5 below.

Update and query. We now illustrate how the update and query operations work in this compressed Fenwick tree. Consider the update operation at the point $(7, 6, 2)$ with value v_3. For each x, y and z level Fenwick tree, we will update all nodes whose Fenwick blocks contain the local compressed index of the update coordinate. We first define a function $\mathrm{UpdatePath}(m, a)$ recovering which child Fenwick nodes to update for a Fenwick tree of size m when the local compressed index of the update coordinate is a, i.e.,

$$q \in \mathrm{UpdatePath}(m, a) \quad \Longleftrightarrow \quad a \in \mathtt{I}(q).$$

Algorithm 3 $\mathrm{UpdatePath}(m, a)$

Require: Local Fenwick size m, local compressed index $a \in \{1, \ldots, m\}$
Ensure: All Fenwick nodes whose blocks contain a

```
P ← []
q ← a
while q ≤ m do
    append q to P
    q ← q + lowbit(q)
end while
return P
```

At the top level, the local compressed index of the first coordinate 7 is 3, so the update path is $\mathrm{UpdatePath}(|\mathtt{T.geometry}|, 3) = [3, 4]$. This implies that we update both `T.child`[3] and `T.child`[4] at the remaining-coordinate location $(6, 2)$.

Updating `T.child`[3] is straightforward: the only remaining coordinate point is $(6, 2)$, so the update is immediately done by adding v_3 on the value at the corresponding coordinate. For `T.child`[4], the local compressed index of 6 is 3, so the update path is $\mathrm{UpdatePath}(|\mathtt{T.child[4].geometry}|, 3) = [3, 4]$. This implies that we update both `T.child`[4]`.child`[3] and `T.child`[4]`.child`[4] at the remaining-coordinate location 2. Note that `T.child`[4]`.child`[3] stores only one point $(6, 2)$, so we directly add v_3 at this point without building a one-dimensional Fenwick tree. On the other hand, `T.child`[4]`.child`[4] is a one-dimensional Fenwick tree with z-coordinates $\{2, 5, 7, 9\}$, so we update this Fenwick tree at the local compressed index of 2, which is 1. Since $\mathrm{UpdatePath}(|\mathtt{T.child[4].child[4].geometry}|, 1) = [1, 2, 4]$,

we update the Fenwick tree at indices $1, 2, 4$ by adding v_3 on the existing values with at these indices. We provide a formal algorithm for the update operation in Algorithm 6 below.

Now, we illustrate how the query operation works in this compressed Fenwick tree. Consider the query `T.query`$(12, 6, 6)$, which asks for the sum of all inserted values at locations satisfying $x \leq 12, y \leq 6, z \leq 6$. Among the five update locations `T.geometry`, the qualifying points are $(7, 6, 2)$ and $(12, 5, 4)$. Therefore the correct answer is `T.query`$(12, 6, 6) = v_3 + v_5$. Let us see how the compressed Fenwick tree obtains this value. First, we define a function $\mathrm{QueryPath}(a)$ which recovers which child Fenwick nodes to query for a Fenwick tree of size m when the query boundary is a, i.e.,

$$\{1, \ldots, a\} = \bigsqcup_{q \in \mathrm{QueryPath}(m,a)} \mathtt{I}(q).$$

Algorithm 4 $\mathrm{QueryPath}(m, a)$

Require: Local prefix index $a \in \{0, \ldots, m\}$; $m \in \mathbb{N}$
Ensure: Fenwick nodes whose blocks form a disjoint partition of $\{1, \ldots, a\}$, or $\emptyset$ if $a = 0$.

```
P ← []
q ← a
while q > 0 do
    append q to P
    q ← q − lowbit(q)
end while
return P
```

Since `T.coords` $= (2, 4, 7, 9, 12)$, the query boundary $x \leq 12$ corresponds to local compressed index 5. Since $\mathrm{QueryPath}(5, 5) = [5, 4]$, the query visits `T.child`[5] and `T.child`[4]. The node `T.child`[5] stores only the remaining coordinate $(5, 4)$, corresponding to the original point $(12, 5, 4)$. Since $(5, 4) \preceq (6, 6)$, this node contributes v_5. Next, `T.child`[4] is queried with boundary $(y, z) \preceq (6, 6)$. Inside `T.child`[4], the y-coordinate list is $\{1, 3, 6, 8\}$. Since the query boundary $y \leq 6$ corresponds to local compressed index 3, and $\mathrm{QueryPath}(4, 3) = [3, 2]$, the query visits `T.child`[4]`.child`[3] and `T.child`[4]`.child`[2]. The `T.child`[4]`.child`[3] stores z-coordinate set $\{2\}$, which contributes v_3 because $2 \leq 6$. The node `T.child`[4]`.child`[2] stores z-coordinate set $\{7, 9\}$, which contributes nothing because both 7 and 9 are larger than 6. This recovers the correct query answer $v_3 + v_5$. We provide a formal algorithm for the query operation in Algorithm 7 below.

Time and memory complexity. Note that the 1-dimensional Fenwick tree supports both query and update in $O(\log m)$ time, where m is the number of coordinates. For k-dimensional nested Fenwick tree we build above, $\mathrm{QueryPath}$ and $\mathrm{UpdatePath}$ output $O(\log m)$ Fenwick nodes at each level, and there are k levels. Since the final level is a one-dimensional Fenwick tree, the query and update time complexity of the k-dimensional compressed Fenwick tree is $O(\log^k m)$, from $O(\log^{k-1} m)$ visits to the final level and $O(\log m)$ time complexity at the final level. For memory complexity, A single update location is copied along $O(\log m)$ Fenwick update-path nodes in each of the first $(k-1)$ coordinates. Hence it reaches $O(\log^{k-1} m)$ final one-dimensional Fenwick trees. Across all m possible update locations, the total number of point-copies reaching the final level is therefore $O(m \log^{k-1} m)$. Since each final one-dimensional compressed Fenwick tree stores only a linear number of slots in the number of locations assigned to it, the total memory usage is $O(m \log^{k-1} m)$.

By applying the dimension-saving trick in Section D.3, we save one dimension in the Fenwick tree, which implies that $k = d - 1$ in the above analysis. Also, our data size is N. Therefore, for each 2^{d-1} orthants and corresponding Fenwick tree, the query and update time complexity of our algorithm is $O(\log^{d-1} N)$, and the memory usage is $O(N \log^{d-2} N)$. This yields total time complexity $O(2^{d-1} N \log^{d-1} N)$ when $d \geq 1$ and total memory usage $O(2^{d-1} N \log^{d-2} N)$ for the entire algorithm when $d \geq 2$. When $d = 1$, by sorting within $O(N \log N)$ time, the kernel lower triangular sum becomes

$$K_{ZZ} = \sum_{j<i} k(z_i, z_j) = f_-(z_i) \sum_{j<i} g_-(z_j),$$

which can be computed in $O(N)$ time and $O(N)$ memory by maintaining the running sum $\sum_{j<i} g_-(z_j)$.

In summary, the total time complexity is $O(2^{d-1} N \log^{(d-1)\vee 1} N)$ and the total memory usage is $O(2^{d-1} N \log^{(d-2)\vee 0} N)$.

Algorithm 5 BuildTree($\mathcal{P}, k$)

Require: Possible update locations $\mathcal{P} \subseteq \mathbb{R}^k$
Ensure: A compressed k-dimensional Fenwick tree node `T`

```
T.k ← k
T.geometry ← P
if |T.geometry| = 1 then
    T.type ← SINGLETON
    T.point ← the unique point in T.geometry
    T.value ← 0
    return T
end if
T.type ← FENWICK
T.coords ← SortUnique{x_1 : (x_1, ..., x_k) ∈ T.geometry}
T.m ← |T.coords|
if k = 1 then
    for q = 1, ..., T.m do
        T.bit[q] ← 0
    end for
    return T
end if
for q = 1, ..., T.m do
    Bucket[q] ← []
end for
for each p = (x_1, ..., x_k) ∈ T.geometry do
    a ← 1 + #{c ∈ T.coords : c < x_1}
    tail ← (x_2, ..., x_k)
    for each q ∈ UpdatePath(T.m, a) do
        append tail to Bucket[q]
    end for
end for
for q = 1, ..., T.m do
    if Bucket[q] ≠ ∅ then
        T.child[q] ← BuildTree(Bucket[q], k − 1)
    else
        T.child[q] ← ∅
    end if
end for
return T
```

Algorithm 6 `T.update`(p, v)

Require: A compressed Fenwick tree node `T`, point $p = (x_1, \ldots, x_k)$, value v
Ensure: Add v at location p

```
if T.type = SINGLETON then
    assert p = T.point
    T.value ← T.value + v
    return
end if
a ← 1 + #{c ∈ T.coords : c < x_1}
if T.k = 1 then
    for each q ∈ UpdatePath(T.m, a) do
        T.bit[q] ← T.bit[q] + v
    end for
    return
end if
tail ← (x_2, ..., x_k)
for each q ∈ UpdatePath(T.m, a) do
    assert T.child[q] ≠ ∅
    T.child[q].update(tail, v)
end for
return
```

Algorithm 7 `T.query`(b)

Require: A compressed Fenwick tree node `T`, boundary $b = (b_1, \ldots, b_k)$
Ensure: Sum of all values at points p satisfying $p \preceq b$

```
if T.type = SINGLETON then
    if T.point ⪯ b then
        return T.value
    else
        return 0
    end if
end if
a ← #{c ∈ T.coords : c ≤ b_1}
ans ← 0
if a = 0 then
    return 0
end if
if T.k = 1 then
    for each q ∈ QueryPath(T.m, a) do
        ans ← ans + T.bit[q]
    end for
    return ans
end if
tailBoundary ← (b_2, ..., b_k)
for each q ∈ QueryPath(T.m, a) do
    if T.child[q] ≠ ∅ then
        ans ← ans + T.child[q].query(tailBoundary)
    end if
end for
return ans
```

Correctness. We now verify that the compressed tree gives the same answer as a standard dense multi-dimensional Fenwick tree restricted to the possible update locations $\texttt{T.geometry}$.

Let $V(p)$ denote the total value inserted so far at location $p \in \texttt{T.geometry}$. We prove the following invariant by induction on the dimension k:

$$\texttt{T.query}(b) = \sum_{\substack{p \in \texttt{T.geometry} \\ p \preceq b}} V(p).$$

For $k = 1$, this is exactly the standard one-dimensional Fenwick tree invariant. Each Fenwick entry q stores the partial sum over the block $\texttt{I}(q)$. An update at local index a adds v to precisely those entries q satisfying $a \in \texttt{I}(q)$, namely $q \in \mathrm{UpdatePath}(\texttt{T.m}, a)$. A query at local prefix index a sums entries over $q \in \mathrm{QueryPath}(\texttt{T.m}, a)$, whose blocks form a disjoint partition of $\{1, \ldots, a\}$. Hence the returned value is the desired prefix sum.

Now suppose the invariant holds for dimension $k - 1$. Consider a k-dimensional compressed tree $\texttt{T}$. By construction, $\texttt{T.child}[q]$ is built from the tails of exactly those update locations whose first-coordinate local rank lies in $\texttt{I}(q)$:

$$\texttt{T.child}[q]\texttt{.geometry} = \big\{ \mathrm{tail}(p) : p = (x_1, \ldots, x_k) \in \texttt{T.geometry},\ R_{\texttt{T}}(x_1) \in \texttt{I}(q) \big\}.$$

When $\texttt{T.update}(p, v)$ is called with first-coordinate local rank $a = R_{\texttt{T}}(x_1)$, the algorithm recursively updates exactly those children q for which

$$q \in \mathrm{UpdatePath}(\texttt{T.m}, a),$$

or equivalently,

$$a \in \texttt{I}(q).$$

Thus every child tree $\texttt{T.child}[q]$ receives exactly the contributions from the update locations whose first-coordinate ranks belong to its Fenwick block $\texttt{I}(q)$.

Next consider a query boundary $b = (b_1, \ldots, b_k)$, and let

$$a = B_{\texttt{T}}(b_1) = \#\{c \in \texttt{T.coords} : c \le b_1\}.$$

The first-coordinate prefix $\{1, \ldots, a\}$ decomposes as the disjoint union

$$\{1, \ldots, a\} = \bigsqcup_{q \in \mathrm{QueryPath}(\texttt{T.m}, a)} \texttt{I}(q).$$

Therefore,

$$\begin{aligned}
\texttt{T.query}(b) &= \sum_{q \in \mathrm{QueryPath}(\texttt{T.m}, a)} \texttt{T.child}[q]\texttt{.query}(b_2, \ldots, b_k) \\
&= \sum_{q \in \mathrm{QueryPath}(\texttt{T.m}, a)} \sum_{\substack{p \in \texttt{T.geometry} \\ R_{\texttt{T}}(p_1) \in \texttt{I}(q) \\ \mathrm{tail}(p) \preceq (b_2, \ldots, b_k)}} V(p),
\end{aligned}$$

where the second equality follows from the induction hypothesis applied to each child tree. Since the blocks $\texttt{I}(q)$ in the query path are disjoint and their union is $\{1, \ldots, a\}$, the last display equals

$$\sum_{\substack{p \in \texttt{T.geometry} \\ R_{\texttt{T}}(p_1) \le a \\ \mathrm{tail}(p) \preceq (b_2, \ldots, b_k)}} V(p).$$

Finally, $R_{\texttt{T}}(p_1) \le a$ is equivalent to $p_1 \le b_1$ among the compressed coordinates. Hence

$$\texttt{T.query}(b) = \sum_{\substack{p \in \texttt{T.geometry} \\ p \preceq b}} V(p),$$

which proves the invariant for dimension k.

Algorithm 8 Sample-splitting PCA-MMD test

Require: Samples $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_m)$; projection dimension r; kernel κ_λ; level α; permutations B.
Ensure: Test decision $\Delta_\alpha^{\mathrm{PCA}}$.
Randomly split $\mathbf{X} = \mathbf{X}^{\mathrm{pca}} \sqcup \mathbf{X}^{\mathrm{test}}$ and $\mathbf{Y} = \mathbf{Y}^{\mathrm{pca}} \sqcup \mathbf{Y}^{\mathrm{test}}$.
Let $\mathbf{Z}^{\mathrm{pca}} = \mathbf{X}^{\mathrm{pca}} \cup \mathbf{Y}^{\mathrm{pca}}$ and compute its mean $\bar{\mathbf{Z}}^{\mathrm{pca}}$ and covariance $\widehat{\Sigma}$.
Let $\widehat{\mathbf{M}} \in \mathbb{R}^{d \times r}$ be the top k eigenvectors of $\widehat{\Sigma}$.
Project the held-out data:

$$\widetilde{\mathbf{X}} = \{\widehat{\mathbf{M}}^\top (x - \bar{Z}^{\mathrm{pca}}) : x \in X^{\mathrm{test}}\}, \qquad \widetilde{\mathbf{Y}} = \{\widehat{\mathbf{M}}^\top (y - \bar{Z}^{\mathrm{pca}}) : y \in Y^{\mathrm{test}}\}.$$

Compute

$$T_0 = \widehat{\mathrm{MMD}}^2(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}; \kappa_\lambda),$$

where all kernel sums in $\widehat{\mathrm{MMD}}^2$ are computed using Algorithm 1.
for $b = 1, \ldots, B$ **do**
  Randomly permute the labels of the pooled projected sample $\widetilde{\mathbf{Z}} = \widetilde{\mathbf{X}} \cup \widetilde{\mathbf{Y}}$ to obtain $(\widetilde{\mathbf{X}}^{\pi_b}, \widetilde{\mathbf{Y}}^{\pi_b})$.
  Compute $T_b = \widehat{\mathrm{MMD}}^2(\widetilde{\mathbf{X}}^{\pi_b}, \widetilde{\mathbf{Y}}^{\pi_b}; \kappa_\lambda)$, again using Algorithm 1.
end for
Let $q_{1-\alpha,B}$ be the $\lceil (1-\alpha)(B+1) \rceil$-th order statistic of $\{T_0, T_1, \ldots, T_B\}$.
return $\Delta_\alpha^{\mathrm{PCA}} = \mathbf{1}\{T_0 > q_{1-\alpha,B}\}$.

E. More details for dimensionality reduction

E.1. Algorithm pseudocode

E.2. Theoretical results

Suppose U, V, ξ, and ξ' have densities μ_U, μ_V, f_ξ and $f_{\xi'}$ respectively. For simiplicity we are assuming $f_\xi = f_{\xi'}$. However, this is not necessary.

Let $\|\nabla f_\xi\|_{L^2(A)} := \left(\int_A \|\nabla f_\xi(z)\|_2^2 \, dz\right)^{1/2}$.

Assumption E.1. Assume that the density f_ξ of ξ satisfies the following:

1. f_ξ is supported on a convex set $A \subseteq \mathbb{R}^d$ and $\ell_2(\mathbf{M}^T\xi, \mathbf{N}^T\xi) \le C\|\mathbf{M} - \mathbf{N}\|$.
2. For some $\ell, L > 0$, $\forall \mathbf{M} \in \mathbb{R}^{d \times r}$ with $\mathbf{M}^T\mathbf{M} = \mathbf{I}_k$

$$\|f_{\mathbf{M}^T\xi}\|_{L_\infty(A)} \le \ell^r \qquad \min\left(\left\|\nabla f_{\mathbf{M}^T\xi}\right\|_{L^2(A)}^2, \left\|\nabla f_{\mathbf{M}^T\xi}\right\|_{L^\infty(A)}^2 \mathrm{Vol}(A)\right) \le L^{2r}.$$

3. The perimeter $\mathrm{Per}(A)$ of A is finite (where $\mathrm{Per}(\mathbb{R}^d) = 0$).

Noise distribution	Density $f(x)$	$\Theta(\|\nabla f\|_{L_p})$
Gaussian	$(2\pi\sigma^2)^{-d/2} e^{-\|x\|_2^2/(2\sigma^2)}$	$\sqrt{1+d/p}\,(\sqrt{2\pi})^{-d(1-1/p)}\sigma^{-1-d(1-1/p)}p^{-\frac{d}{2p}}$
Product Laplace	$(2b)^{-d} e^{-\|x\|_1/b}$	$\sqrt{d}\,2^{-d(1-1/p)} b^{-1-d(1-1/p)} p^{-d/p}$
Uniform on A	$\mathrm{Vol}(A)^{-1}\mathbf{1}_A(x)$	∞

Table 3. L_p gradient norms for common noise distributions up to constant factors.

Lemma E.2. *Suppose the density f_ξ is supported on a convex $A \subseteq \mathbb{R}^d$ and differentiable in the interior of A. We have:*

$$\int_{\mathbb{R}^d} (f_\xi(Z+h) - f_\xi(Z))^2 \le \|h\|_2^2 \min\left(\|\nabla f_\xi\|_{L^2(A)}^2, \|\nabla f_\xi\|_{L^\infty(A)}^2 \mathrm{Vol}(A)\right) + 2\|f_\xi\|_{L_\infty(A)}^2 \mathrm{Per}(A)\|h\|_2.$$

Proof. For any vector $h \in \mathbb{R}^d$, define the good set

$$G_h = \{Z \in A : Z + h \in A\}.$$

For any $Z \in G_h$, the line segment joining Z and $Z + h$ stays inside A because A is convex, so

$$f_\xi(Z + h) - f_\xi(Z) = \int_0^1 \nabla f_\xi(Z + th)^T h \, dt.$$

It follows that

$$\begin{aligned}
\int_{G_h} (f_\xi(Z + h) - f_\xi(Z))^2 \, dZ &= \int_{G_h} \left(\int_0^1 \nabla f_\xi(Z + th)^T h \, dt \right)^2 dZ \\
&\leq \int_{G_h} \left(\int_0^1 \left(\nabla f_\xi(Z + th)^T h \right)_2^2 dt \int_0^1 1^2 \, dt \right) dZ \\
&\leq \|h\|_2^2 \int_{G_h} \int_0^1 \left\| \nabla f_\xi(Z + th)^T \right\|_2^2 dt dZ \\
&= \|h\|_2^2 \int_0^1 \int_{G_h} \left\| \nabla f_\xi(Z + th)^T \right\|_2^2 dt dZ \leq \|h\|_2^2 \, \|\nabla f_\xi\|_{L^2(A)}^2 \\
&\leq \|h\|_2^2 \min \left(\|\nabla f_\xi\|_{L^2(A)}^2, \|\nabla f_\xi\|_{L^\infty(A)}^2 \operatorname{Vol}(A) \right).
\end{aligned}$$

Next, define the bad set

$$B_h := \{Z : \text{ exactly one of } Z \text{ and } Z + h \text{ is in } A\}.$$

On this set,

$$\int_{B_h} (f_\xi(Z + h) - f_\xi(Z))^2 dZ = \int_{B_h} (f_\xi(Z)^2 + f_\xi(Z + h)^2) dZ \leq 2\|f_\xi\|_{L^\infty(A)}^2 \operatorname{Vol}(B_h)$$

For convex bodies A, $\operatorname{Vol}(B_h) \leq \operatorname{Per}(A) \, \|h\|_2$, so

$$\int_{B_h} (f_\xi(Z + h) - f_\xi(Z))^2 dZ \leq 2 \, \|f_\xi\|_{L^\infty(A)}^2 \operatorname{Per}(A) \, \|h\|_2 .$$

Therefore,

$$\int_{\mathbb{R}^d} (f_\xi(Z + h) - f_\xi(Z))^2 \leq \|h\|_2^2 \min \left(\|\nabla f_\xi\|_{L^2(A)}^2, \|\nabla f_\xi\|_{L^\infty(A)}^2 \operatorname{Vol}(A) \right) + 2 \, \|f_\xi\|_{L^\infty(A)}^2 \operatorname{Per}(A) \, \|h\|_2 .$$

□

Lemma E.3. *Define $\ell_2(\mu, \nu)$ be the ℓ_2 distance between two densitites μ and ν. Assume that the common density f of ξ and ξ' satisfy assumption E.1. Let $\max(\mathbb{E}[\|U\|^2], \mathbb{E}[\|V\|^2]) \leq m_r^2$. Then, for any orthonormal $\widehat{\mathbf{M}}$ such that $\|\widehat{\mathbf{M}} - \mathbf{MO}\| \leq \epsilon$ for some rotation matrix $\mathbf{O}$,*

$$\ell_2(\widehat{\mathbf{M}}^T X, \widehat{\mathbf{M}}^T Y) \geq \ell_2(\mathbf{O}^T \mathbf{M}^T X, \mathbf{O}^T \mathbf{M}^T Y) - 3\sqrt{L^{2r} \epsilon^2 m_r^2 + 2\ell^r \epsilon m_r \mathit{Per}(A) + 2C^2 \epsilon^2 m_r^2}$$

Proof. First, note that, using triangle inequality,

$$\ell_2(\widehat{\mathbf{M}}^T X, \widehat{\mathbf{M}}^T Y) \geq \ell_2(\mathbf{O}^T \mathbf{M}^T X, \mathbf{O}^T \mathbf{M}^T Y) - \ell_2(\mathbf{O}^T \mathbf{M}^T X, \widehat{\mathbf{M}}^T X) - \ell_2(\mathbf{O}^T \mathbf{M}^T Y, \widehat{\mathbf{M}}^T Y) \tag{25}$$

Now we will show the two second terms in the RHS of Eq 25 can be upper bounded. We have:

$$\widehat{\mathbf{M}}^T X = \mathbf{O}^T \mathbf{M}^T X + (\widehat{\mathbf{M}} - \mathbf{MO})^T X = \mathbf{O}^T U + \mathbf{O}^T \mathbf{M}^T \xi + (\widehat{\mathbf{M}} - \mathbf{MO})^T X$$

First, note that:

$$f_{\mathbf{M}^T X}(Z) = \int_U f_{\mathbf{U}}(U) f_{\mathbf{M}^T \xi}(Z-U) dU = \mathbb{E}_U[f_{\mathbf{M}^T\xi}(Z-U)]$$

$$\begin{aligned}
&\ell_2\left(\mu_{\mathbf{O}^T\mathbf{M}^T X}, \mu_{\widehat{\mathbf{M}}^T X}\right)^2 = \int_Z \left(\mathbb{E}_U[f_{\mathbf{O}^T\mathbf{M}^T\xi}(Z-\mathbf{O}^T U)] - \mathbb{E}_U[f_{\widehat{\mathbf{M}}^T\xi}(Z-\widehat{\mathbf{M}}^T\mathbf{M}U)]\right)^2 dZ \\
&\leq \int_Z \mathbb{E}_U\left(f_{\mathbf{O}^T\mathbf{M}^T\xi}(Z-\mathbf{O}^T U) - f_{\widehat{\mathbf{M}}^T\xi}(Z-\widehat{\mathbf{M}}^T\mathbf{M}U)\right)^2 dZ \\
&\leq 2\mathbb{E}_U \int_Z \left(f_{\mathbf{O}^T\mathbf{M}^T\xi}(Z-\mathbf{O}^T U) - f_{\mathbf{O}^T\mathbf{M}^T\xi}(Z-\widehat{\mathbf{M}}^T\mathbf{M}U)\right)^2 dZ \\
&+ 2\mathbb{E}_U \int_Z \left(f_{\mathbf{O}^T\mathbf{M}^T\xi}(Z-\widehat{\mathbf{M}}^T\mathbf{M}U) - f_{\widehat{\mathbf{M}}^T\xi}(Z-\widehat{\mathbf{M}}^T\mathbf{M}U)\right)^2 dZ \\
&\leq 2L^{2r}\|\mathbf{O}-\widehat{\mathbf{M}}^T\mathbf{M}\|^2\mathbb{E}[\|U\|^2] + 2\ell^r \mathrm{Per}(A)\|\mathbf{O}-\widehat{\mathbf{M}}^T\mathbf{M}\|\mathbb{E}[\|U\|] + 2C^2\|\mathbf{O}^T\mathbf{M}^T - \widehat{\mathbf{M}}\|^2\mathbb{E}[\|U\|^2]
\end{aligned}$$

This follows from Lemma E.2. Using an analogous bound for $\mathbf{M}^T Y$, we get the final bound. □

Furthermore, assume that the following regularity condition holds on the density of the noise. Define $\left\|\nabla f_{\mathbf{M}^T\xi}\right\|_{L_2}^2 = \int_Z \left\|\nabla f_{\mathbf{M}^T\xi}(Z)\right\|^2 dZ$.

Corollary E.4. *Define $\ell_2(\mu,\nu)$ be the ℓ_2 distance between two densitites μ and ν. Assume that the common density f_ξ and $f_{\xi'}$ satisfy assumption E.1 and have full support. Let $\max(\mathbb{E}[\|U\|^2], \mathbb{E}[\|V\|^2]) \leq m_r^2$. Then, for any orthonormal $\widehat{\mathbf{M}}$ such that $\|\widehat{\mathbf{M}} - \mathbf{M}\mathbf{O}\| \leq \epsilon$ for some rotation matrix $\mathbf{O}$,*

$$\ell_2(\widehat{\mathbf{M}}^T X, \widehat{\mathbf{M}}^T Y) \geq \ell_2(\mathbf{O}^T\mathbf{M}^T X, \mathbf{O}^T\mathbf{M}^T Y) - 3\epsilon m_r\sqrt{L^{2r}+C^2}$$

Proof of Corollary E.4. This follows directly from Lemma E.3 with $A = \mathbb{R}^d$. □

Verification of Assumption E.1 for isotropic Gaussian noise. Suppose $\xi \sim N(0, \sigma^2 I_d)$. FOR any $M \in \mathbb{R}^{d\times r}$ satisfying $M^\top M = \mathbf{I}_k$,

$$M^\top \xi \sim N(0, \sigma^2 \mathbf{I}_k).$$

Hence

$$f_{M^\top\xi}(z) = (2\pi\sigma^2)^{-r/2}\exp\left(-\frac{\|z\|_2^2}{2\sigma^2}\right), \qquad z \in \mathbb{R}^r.$$

Therefore

$$\|f_{M^\top\xi}\|_\infty = (2\pi\sigma^2)^{-r/2} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^r.$$

Moreover,

$$\nabla f_{M^\top\xi}(z) = -\frac{z}{\sigma^2} f_{M^\top\xi}(z),$$

and hence

$$\begin{aligned}
\|\nabla f_{M^\top\xi}\|_{L_2}^2 &= \int_{\mathbb{R}^r} \frac{\|z\|_2^2}{\sigma^4} f_{M^\top\xi}(z)^2\, dz \\
&= \frac{r}{2\sigma^2}(4\pi\sigma^2)^{-r/2} = \frac{r}{2\sigma^2}\left(\frac{1}{2\sqrt{\pi}\sigma}\right)^r.
\end{aligned}$$

Thus Assumption 1 holds with

$$\ell = \frac{1}{\sqrt{2\pi}\sigma}, \qquad L = \frac{1}{2\sqrt{\pi}\sigma}\left(\frac{r}{2\sigma^2}\right)^{1/r}.$$

Finally, for any two orthonormal matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times r}$, both $\mathbf{M}^\top\xi$ and $\mathbf{N}^\top\xi$ have law $\mathbf{N}(0, \sigma^2\mathbf{I}_k)$. Hence

$$\ell_2(\mathbf{M}^\top\xi, \mathbf{N}^\top\xi) = 0,$$

so the Lipschitz condition

$$\ell_2(\mathbf{M}^\top \xi, \mathbf{N}^\top \xi) \le C\|\mathbf{M} - \mathbf{N}\|$$

holds with $C = 0$.

Proof of Lemma 5.3

Proof. Some calculation (see above) shows that: $L = \frac{1}{2\sqrt{\pi}\sigma}\left(\frac{r}{2\sigma^2}\right)^{1/r}$.

Let the pooled covariance matrix be $\hat{\Sigma} = \frac{1}{N}\sum_i Z_i Z_i^T - \bar{Z}\bar{Z}^T$. We have

$$\Sigma := \mathbb{E}[\hat{\Sigma}] = \alpha \mathbf{M}\Sigma_U \mathbf{M}^T + (1-\alpha)\mathbf{M}\Sigma_V \mathbf{M}^T + \mathbf{I}_r,$$

where $\alpha = n/N$, Σ_U (Σ_V) denotes the covariance matrix of U (V). We see that $\lambda_{k+1}(\Sigma) = 1$ and $\lambda_r(\Sigma) \ge 1 + \lambda_r(\alpha\Sigma_U + (1-\alpha)\Sigma_V)$. Furthermore we have, $\lambda_r(\alpha\Sigma_U + (1-\alpha)\Sigma_V) \ge \gamma/2$, by definition. Now using Theorem 2 in (Yu et al., 2015), we have:

$$\|\hat{\mathbf{M}} - \mathbf{M}\mathbf{O}\| \le \frac{2\sqrt{8r}\|\hat{\Sigma} - \Sigma\|}{\gamma}$$

Note that the pooled covariance matrix is an average of $\hat{\Sigma}_1 = \sum_{i=1}^n X_i X_i^T/n$ and $\hat{\Sigma}_2 = \sum_{i=1}^m Y_i Y_i^T/m$. Using Corollary 1.2 in (Srivastava & Vershynin, 2013), we see under the strong regularity condition on the whitened vector $\Sigma_U^{-1/2}U$ and $\Sigma_V^{-1/2}V$, there exists a constant $C_{\mathrm{cov}} > 0$, depending only on the strong-regularity constants C_{SR} and η, such that the sample covariance matrix $\Sigma_N = N^{-1}\sum_{i=1}^N X_i X_i^\top$ satisfies

$$\begin{aligned}
\mathbb{E}\|\widehat{\Sigma} - \Sigma\| &\le \alpha\,\mathbb{E}\|\widehat{\Sigma}_1 - (\mathbf{M}\Sigma_u\mathbf{M}^\top + I_d)\| + (1-\alpha)\,\mathbb{E}\|\widehat{\Sigma}_2 - (\mathbf{M}\Sigma_v\mathbf{M}^\top + I_d)\| \\
&\le C_{\mathrm{cov}}\alpha\|\mathbf{M}\Sigma_u\mathbf{M}^\top + I_d\|\left(\frac{d}{n}\right)^{\eta/(2\eta+2)} + C_{\mathrm{cov}}(1-\alpha)\|\mathbf{M}\Sigma_V\mathbf{M}^\top + I_d\|\left(\frac{d}{m}\right)^{\eta/(2\eta+2)} \\
&\le C_{\mathrm{cov}}(1+\lambda)\left(\frac{d}{\min(m,n)}\right)^{\eta/(2\eta+2)},
\end{aligned}$$

$$a := \frac{\eta}{2\eta+2}, \qquad b := \frac{2s}{4s+r}.$$

For Gaussian noise with $\sigma = 1$,

$$L = \frac{1}{2\sqrt{\pi}}\left(\frac{r}{2}\right)^{1/r}, \qquad L^r = \frac{r}{2(2\sqrt{\pi})^r}.$$

When the two sample sizes are balanced, the PCA-induced loss is controlled by $A\left(\frac{d}{n_{\mathrm{pca}} \wedge m_{\mathrm{pca}}}\right)^a$, where

$$A := \frac{C_{\mathrm{cov}}(1+\lambda)}{\gamma} m_r \sqrt{2r}\, L^r = \frac{C_{\mathrm{cov}}(1+\lambda)}{\gamma} m_r \frac{r^{3/2}}{\sqrt{2}(2\sqrt{\pi})^r}.$$

The k-dimensional lower-bound scale is

$$\rho_{n,r} \gtrsim n_{\mathrm{test}}^{-b} r^{-rs/\{2(4s+r)\}} R_r^{r/(4s+r)}.$$

Thus the PCA error is dominated by the minimax scale if

$$A\left(\frac{d}{n_{\mathrm{pca}} \wedge m_{\mathrm{pca}}}\right)^a \lesssim n_{\mathrm{test}}^{-b} r^{-rs/\{2(4s+r)\}} R_r^{r/(4s+r)}.$$

With $n_{\mathrm{pca}} = n/2$, $m_{\mathrm{pca}} = m/2$, and $n_{\mathrm{test}} = n/2$, this becomes

$$\frac{n \wedge m}{2} \gtrsim d\left[A\, r^{rs/\{2(4s+r)\}} R_r^{-r/(4s+r)}\right]^{1/a}\left(\frac{n}{2}\right)^{b/a}.$$

Using

$$A \asymp \frac{C_{\mathrm{cov}}(1+\lambda)}{\gamma} m_r \frac{r^{3/2}}{(2\sqrt{\pi})^r},$$

a sufficient condition is

$$n \wedge m \gtrsim d \left[\frac{C_{\mathrm{cov}}(1+\lambda)}{\gamma} m_r \frac{r^{3/2}}{(2\sqrt{\pi})^r} r^{rs/\{2(4s+r)\}} R_r^{-r/(4s+r)} \right]^{(2\eta+2)/\eta} n^{4s(\eta+1)/\{\eta(4s+r)\}}.$$

If the two samples are balanced and $\eta > 4s/r$, this is implied by

$$n \gtrsim \left\{ d \left[\frac{C_{\mathrm{cov}}(1+\lambda)}{\gamma} m_r \frac{r^{3/2}}{(2\sqrt{\pi})^r} r^{rs/\{2(4s+r)\}} R_r^{-r/(4s+r)} \right]^{(2\eta+2)/\eta} \right\}^{\eta(4s+r)/(\eta r-4s)}.$$

The final bound is,

$$n \gtrsim C(r, \eta, C_{\mathrm{cov}}, \lambda/\gamma, m_r) d^{\frac{\eta(4s+r)}{\eta r-4s}}$$

□

F. MNIST experiments

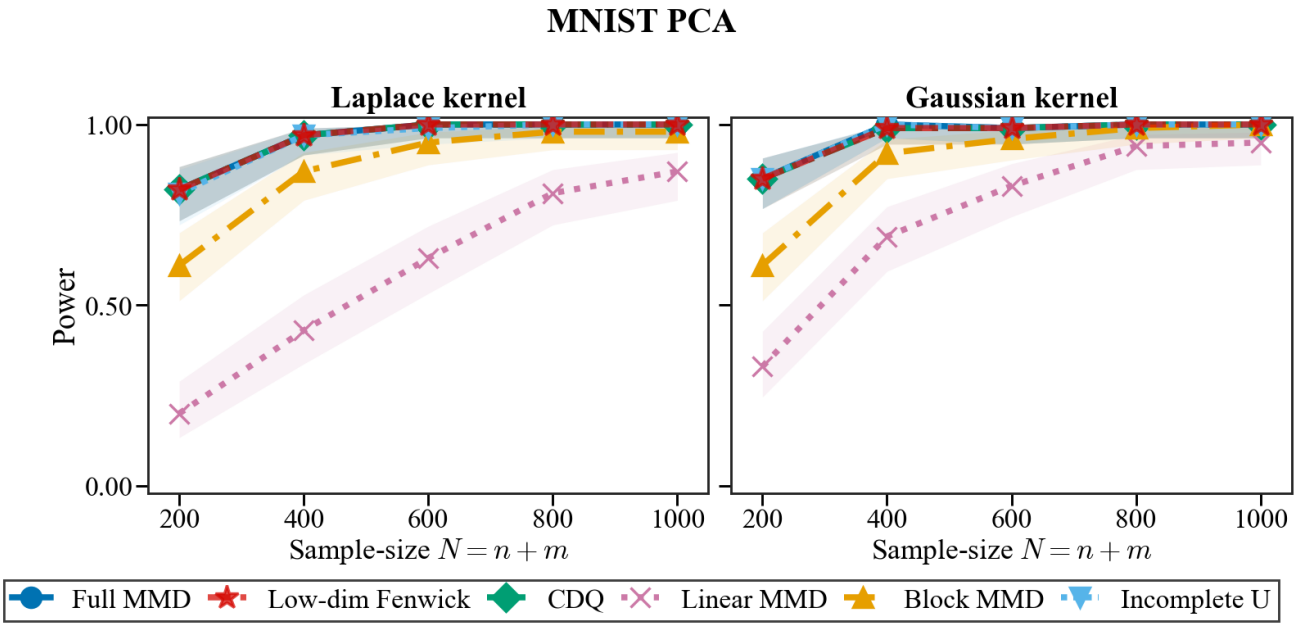

Figure 3. MMD permutation two sample testing with MNIST Datasets, with dimension redcution via PCA.

We also use the MNIST data to do the hypothesis test that decides between the even numbers and odd numbers. For $N \in [200, \ldots, 2000]$, we draw $n = N/2$ samples from the training samples with an even label and $m = N/2$ samples from the training samples with an even label. The signal dimension is $r = 2$. As in PCA-MMD, we estimate the top r principal components from the pooled sample $X \cup Y$, project all samples to this dimension, and run the MMD permutation test on the k-dimensional data. As shown in Figure 3, the proposed low-dimensional Fenwick approach achieves essentially the same power as the full MMD permutation test for both Laplace and Gaussian kernels. This indicates that the Fenwick-based acceleration does not incur a noticeable loss in testing power in this experiment, while providing a computationally more efficient way to evaluate the MMD statistic.