BOOSTING THE POWER OF KERNEL TWO-SAMPLE TESTS

ANIRBAN CHATTERJEE AND BHASWAR B. BHATTACHARYA

ABSTRACT. The kernel two-sample test based on the maximum mean discrepancy (MMD) is one of the most popular methods for detecting differences between two distributions over general metric spaces. In this paper we propose a method to boost the power of the kernel test by combining MMD estimates over multiple kernels using their Mahalanobis distance. We derive the asymptotic null distribution of the proposed test statistic and use a multiplier bootstrap approach to efficiently compute the rejection region. The resulting test is universally consistent and, since it is obtained by aggregating over a collection of kernels/bandwidths, is more powerful in detecting a wide range of alternatives in finite samples. We also derive the distribution of the test statistic for both fixed and local contiguous alternatives. The latter, in particular, implies that the proposed test is statistically efficient, that is, it has non-trivial asymptotic (Pitman) efficiency. Extensive numerical experiments are performed on both synthetic and real-world datasets to illustrate the efficacy of the proposed method over single kernel tests. Our asymptotic results rely on deriving the joint distribution of MMD estimates using the framework of multiple stochastic integrals, which is more broadly useful, specifically, in understanding the efficiency properties of recently proposed adaptive MMD tests based on kernel aggregation.

1. INTRODUCTION

Given two probability distributions P and Q on a separable metric space $\mathcal{X}$, the two-sample problem is to test the hypothesis:

$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q, \tag{1.1}$$

based on i.i.d. samples $\mathscr{X}_m := \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n := \{Y_1, Y_2, \ldots, Y_n\}$ from the distributions P and Q, respectively. This is a classical problem that has been extensively studied, especially in the parametric regime, where the data is assumed to have certain low-dimensional functional forms. However, parametric methods often perform poorly for misspecified models, especially when the number of nuisance parameters is large, and for non-Euclidean data. This necessitates the development of non-parametric methods, which make minimal distributional assumptions on the data, but remain powerful for a wide class of alternatives.

For univariate data, there are several well-known nonparametric tests such as the two-sample Kolmogorov–Smirnoff (KS) maximum deviation test [59], the Wald-Wolfwotiz runs test [65], the rank-sum test [46, 67], and the Cramér-von Mises test [1]. Efforts to generalize these univariate methods to higher dimensions date back to Weiss [66] and Bickel [9]. Thereafter, several nonparametric methods for multivariate two-sample testing have been proposed over the years. These include tests based on geometric graphs [7, 10, 11, 19, 25–28, 52–54], tests based on data-depth [42, 43], the energy distance test [3, 5, 61–63], kernel maximum mean discrepancy (MMD) tests [13, 22–24, 48, 49, 56, 58, 60, 68], ball divergence [4, 47], projection-averaging [33], classifier-based tests [34, 44], among others. Recently, distribution-free versions of the energy distance/kernel tests have been proposed by Deb and Sen [14] and Deb et al. [15], using the emerging theory of multivariate ranks based on optimal transport.

Key words and phrases. Kernel methods, nonparametric two-sample testing, Pitman efficiency, U-statistics.
The research was partly supported by NSF CAREER Grant DMS-2046393 and a Sloan research fellowship.

Among the aforementioned methods kernel-based tests have emerged as a powerful technique for detecting distributional differences on general domains. The basic idea is to quantify the discrepancy between the two distributions P and Q in terms of the largest difference in expectation between $f(X)$ and $f(Y)$, for $X \sim P$ and $Y \sim Q$, over functions f in the unit ball of a reproducing kernel Hilbert space (RKHS) defined on $\mathcal{X}$. This is called the *maximum mean discrepancy* (MMD) between the distributions P and Q (see (2.1) for the precise definition), which can be conveniently estimated from the data in terms of the pairwise kernel dissimilarities (see Section 2.1 for details). Another useful property of the MMD is that it takes value zero if and only if the distributions P and Q are the same. Consequently, the test which rejects H_0 for large values of the estimated MMD is universally consistent (the power of the test converges to 1 as the sample size increases) for the hypothesis (1.1) (see Gretton et al. [23] for further details).

Although the kernel two-sample test is widely used and has found numerous applications, it often performs poorly for high-dimensional problems [48] and its empirical performance depends heavily on the choice of the kernel. Kernels are usually parametrized by their bandwidths, and the most popular strategy for choosing the kernel bandwidth is the *median heuristics*, where the bandwidth is chosen to be the median of the pairwise distances of the pooled sample [23]. Despite its popularity there is limited understanding of the median heuristic and empirical results demonstrate that the median heuristic performs poorly when differences between the 2 distributions occur at a scale that differs significantly from the median of the interpoint distances [24]. Another approach is to split the data and estimate the kernel by maximizing an approximate empirical power on the held-out data [24, 41]. This, however, can lead to loss in power for smaller sample sizes. (For a discussion of other related methods see Section 9.)

In this paper we propose a strategy for augmenting the power of the classical (single) kernel two-sample test by borrowing strengths from multiple kernels. Specifically, we propose a new test statistic which combines MMD estimates from $r \geqslant 1$ kernels using their sample Mahalanobis distance. The advantage of aggregating across a collection of kernels/bandwidths is that the test can simultaneously deal with cases which require both small and large bandwidths, and, hence, detect both global and local differences more effectively. To understand the asymptotic properties of the test we derive the joint distribution of the vector of MMD estimates under H_0, which can be described using bivariate stochastic integrals, and, as a consequence, derive the asymptotic distribution of the Mahalanobis aggregated MMD (MMMD) statistic under H_0 (Section 3). Moreover, using the kernel Gram matrix representation we develop a multiplier bootstrap approach that allows us to efficiently compute the rejection threshold for the MMMD statistic and show that the resulting test is universally consistent (Section 4). Next, we establish the asymptotic (Pitman) efficiency of the proposed test by deriving its power against local alternatives in the well-known contamination model (Section 5). In Section 6 we derive the joint distribution of MMD estimates and, consequently, that of the MMMD statistic, under the alternative.

In Section 7 we perform extensive simulations to compare our MMMD based test with various single kernel MMD tests (with bandwidths chosen based on the median heuristic). The experiments show that the MMMD method outperforms the single kernel tests and also the graph-based Friedman-Rafsky test [19] across a range alternatives and dimensions, showcasing the efficacy of our aggregation method. In Section 8 we apply the proposed method to compare images of digits in the noisy MNIST dataset. The MMMD effectively distinguishes different digits for significantly more noisy images compared to its single kernel counterparts, again illustrating the advantage of using multiple kernels.

Our results on the joint distribution for multiple kernels are also more broadly useful in understanding the asymptotic properties of general aggregation strategies. To demonstrate

this, in Section 9, we propose an asymptotic implementation of the adaptive MMD test recently proposed in [55], and derive its asymptotic local power. Numerical results comparing the MMMD method and the aforementioned adaptive test are also reported. The codes for all the experiments can be found in the Github repository https://github.com/anirbanc96/MMMD-boost-kernel-two-sample.

2. Kernel Maximum Mean Discrepancy and Mahalanobis Aggregation

We begin by recalling the fundamentals of the kernel two-sample test as introduced in Gretton et al. [23] in Section 2.1. Then in Section 2.2 we describe our proposed test statistic obtained by combining multiple kernels.

2.1. Kernel Maximum Mean Discrepancy.

Suppose $\mathcal{X}$ is a separable metric space and $\mathscr{B}(\mathcal{X})$ is the sigma-algebra generated by the open sets of $\mathcal{X}$. Denote by $\mathcal{P}(\mathcal{X})$ the collection of all probability distributions on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$. Suppose $P, Q \in \mathcal{P}(X)$ and $X \sim P$ and $Y \sim Q$ be random variables distributed as P and Q, respectively. Throughout we will assume that P and Q are *non-atomic*. The maximum mean discrepancy (MMD) between P and Q is defined as

$$\mathrm{MMD}\left[\mathcal{F}, P, Q\right] = \sup_{f \in \mathcal{F}} \left\{\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]\right\}, \tag{2.1}$$

where $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ defined on $\mathcal{X}$ [2]. Since $\mathcal{H}$ is an RKHS, for every $x \in \mathcal{X}$ the evaluation map operator $\eta_x : \mathcal{H} \to \mathbb{R}$ given $\eta_x(f) = f(x)$ is continuous. Thus, by the Riesz representation theorem [50, Theorem II.4] for each $x \in \mathcal{X}$ there is a feature mapping $\psi_x \in \mathcal{H}$ such that $f(x) = \langle f, \psi_x \rangle_{\mathcal{H}}$, for every $f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. The feature mapping takes the canonical form $\psi_x(\cdot) = \mathsf{K}(x, \cdot)$, where $\mathsf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel. This, in particular, implies that $\mathsf{K}(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$. Extending the notion of feature map, an element $\mu_P \in \mathcal{H}$ is defined to be the *mean embedding* of $P \in \mathcal{P}(\mathcal{X})$ if

$$\langle f, \mu_P \rangle_{\mathcal{H}} = \mathbb{E}_{X\sim P}[f(X)], \tag{2.2}$$

for all $f \in \mathcal{H}$. By the canonical form of the feature map it follows that

$$\mu_P(t) := \int_{\mathcal{X}} \psi_t(x) \mathrm{d}P(x) = \mathbb{E}_{X\sim P}[\psi_t(X)] = \mathbb{E}_{X\sim P}[\mathsf{K}(t, X)]. \tag{2.3}$$

Throughout we will make the following assumption:

Assumption 2.1. The kernel $\mathsf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies the following:

(1) $\mathbb{E}_{X\sim P}[\mathsf{K}(X, X)^{\frac{1}{2}}] < \infty$ and $\mathbb{E}_{Y\sim Q}[\mathsf{K}(Y, Y)^{\frac{1}{2}}] < \infty$.
(2) K is *characteristic*, that is, the mean embedding $\mu : \mathcal{P}(\mathcal{X}) \to \mathcal{H}$ is a one-to-one (injective) function.

Assumption 2.1 ensures that $\mu_P, \mu_Q \in \mathcal{H}$ and MMD defines a metric on $\mathcal{P}(\mathcal{X})$. Then the MMD can be expressed as the distance between mean embeddings in $\mathcal{H}$ (see [23, Lemma 4]):

$$\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right] = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}, \tag{2.4}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm corresponding to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. This implies $\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right] = 0$ if and only if $P = Q$. Expanding the square in (2.4) and using the representation in (2.3) it follows that (see [23, Lemma 6] for details)

$$\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right] = \mathbb{E}_{X,X'\sim P}[\mathsf{K}(X, X')] + \mathbb{E}_{Y,Y'\sim Q}[\mathsf{K}(Y, Y')] - 2\mathbb{E}_{X\sim P, Y\sim Q}[\mathsf{K}(X, Y)].$$

Therefore, based on i.i.d. observations $\mathscr{X}_m := \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n := \{Y_1, Y_2, \ldots, Y_n\}$, a natural unbiased estimate of $\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]$ is given by,

$$\mathrm{MMD}^2\left[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n\right] = \mathcal{W}_{\mathscr{X}_m} + \mathcal{W}_{\mathscr{Y}_n} - 2\mathcal{B}_{\mathscr{X}_m, \mathscr{Y}_n}, \tag{2.5}$$

where

$$\mathcal{W}_{\mathscr{X}_m} := \frac{1}{m(m-1)} \sum_{1 \leqslant i \neq j \leqslant m} \mathsf{K}\left(X_i, X_j\right) \text{ and } \mathcal{W}_{\mathscr{Y}_n} := \frac{1}{n(n-1)} \sum_{1 \leqslant i \neq j \leqslant n} \mathsf{K}\left(Y_i, Y_j\right) \tag{2.6}$$

is the average of the kernel dissimilarities within the samples in $\mathscr{X}_m$ and $\mathscr{Y}_n$, respectively, and

$$\mathcal{B}_{\mathscr{X}_m, \mathscr{Y}_n} := \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathsf{K}\left(X_i, Y_j\right) \tag{2.7}$$

is the average of the kernel dissimilarities between the samples in $\mathscr{X}_m$ and $\mathscr{Y}_n$. In the asymptotic regime where $N := m + n \to \infty$ such that

$$\frac{m}{m+n} \to \rho \in (0, 1), \tag{2.8}$$

$\mathrm{MMD}^2[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n]$ is a consistent estimate of $\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]$ (see [23, Theorem 7]), that is,

$$\mathrm{MMD}^2[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{P} \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]. \tag{2.9}$$

Hence, the test which rejects H_0 in (1.1) for large values of $\mathrm{MMD}^2\left[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$ is universally consistent.

2.2. **Aggregating Multiple Kernels.** Fix $r \geqslant 1$ and suppose $\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r$ be r distinct kernels each of which satisfy Assumption 2.1. Denote the vector of MMD estimates as

$$\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right] = \left(\mathrm{MMD}^2[\mathsf{K}_1, \mathscr{X}_m, \mathscr{Y}_n], \cdots, \mathrm{MMD}^2[\mathsf{K}_r, \mathscr{X}_m, \mathscr{Y}_n]\right)^\top, \tag{2.10}$$

where $\mathcal{K} := \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$. In this paper we propose a new test statistic that combines the contributions of the different kernels using the Mahalanobis distance of the vector $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$ as follows:

$$\left(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]\right)^\top \boldsymbol{S}^{-1} \left(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]\right), \tag{2.11}$$

where is $\boldsymbol{S}$ is a consistent estimate of the limiting covariance matrix of $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$ under H_0 (which we denote by $\boldsymbol{\Sigma}_{H_0} = ((\sigma_{ab}))_{1 \leqslant a,b \leqslant r}$). Note that adjusting by the covariance matrix $\boldsymbol{S}$ brings the contributions of the individual MMD estimates in the same scale and by selecting a range of kernels/bandwidths in $\mathcal{K}$ one can detect more fine-grained deviations from H_0, leading to significant power improvements as in will be seen in Section 7. (In Appendix E we present general conditions under which $\boldsymbol{\Sigma}_{H_0}$ is invertible, which, in particular, hold for the any collection of Gaussian or Laplace kernels.)

In Corollary 3.1 we compute

$$\begin{aligned} \sigma_{ab} &:= \lim_{N \to \infty} (m+n)^2 \left(\mathrm{Cov}_{H_0}\left[\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]\right]\right)_{ab} \\ &= \frac{2}{\rho^2(1-\rho)^2} \mathbb{E}_{X, X' \sim P}\left[\mathsf{K}_a^\circ(X, X') \mathsf{K}_b^\circ(X, X')\right], \end{aligned} \tag{2.12}$$

where

$$\mathsf{K}_a^\circ(x, y) = \mathsf{K}_a(x, y) - \mathbb{E}_{X \sim P} \mathsf{K}_a(X, y) - \mathbb{E}_{X' \sim P} \mathsf{K}_a(x, X') + \mathbb{E}_{X, X' \sim P} \mathsf{K}_a(X, X') \tag{2.13}$$

is the centered version of the kernel K_a, for $1 \leqslant a \leqslant r$. Therefore, a natural empirical estimate of $\mathbf{\Sigma}_{H_0}$ is the given by $\hat{\mathbf{\Sigma}} = ((\hat{\sigma}_{ab}))_{1\leqslant a,b\leqslant r}$, where

$$\hat{\sigma}_{ab} = \frac{2}{\hat{\rho}^2(1-\hat{\rho})^2} \cdot \frac{1}{m^2} \sum_{1\leqslant i,j\leqslant m} \hat{\mathsf{K}}_a^{\circ}(X_i, X_j)\hat{\mathsf{K}}_b^{\circ}(X_i, X_j), \tag{2.14}$$

with

$$\hat{\mathsf{K}}_a^{\circ}(x,y) = \mathsf{K}_a(x,y) - \frac{1}{m}\sum_{u=1}^{m} \mathsf{K}_a(X_u, y) - \frac{1}{m}\sum_{v=1}^{m} \mathsf{K}_a(x, X_v) + \frac{1}{m^2} \sum_{1\leqslant u,v\leqslant m} \mathsf{K}_a(X_u, X_v) \tag{2.15}$$

being the empirical analogue of K_a° and $\hat{\rho} = \frac{m}{m+n}$. Therefore, choosing $\boldsymbol{S} = \hat{\mathbf{\Sigma}}$ in (2.11) we define the *Mahalanobis aggregated MMD* (MMMD) statistic as follows:

$$T_{m,n} := \left(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]\right)^{\top} \hat{\mathbf{\Sigma}}^{-1} \left(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]\right). \tag{2.16}$$

In Corollary 3.1 we show that $\hat{\mathbf{\Sigma}} \xrightarrow{P} \mathbf{\Sigma}_{H_0}$, hence (2.9) implies that

$$T_{m,n} \xrightarrow{P} \left(\mathrm{MMD}^2\left[\mathcal{K}, P, Q\right]\right)^{\top} \mathbf{\Sigma}_{H_0}^{-1} \left(\mathrm{MMD}^2\left[\mathcal{K}, P, Q\right]\right) := T_{\mathcal{K}}. \tag{2.17}$$

Note that $T_{\mathcal{K}} = 0$ under H_0 and $T_{\mathcal{K}} > 0$ whenever $P \neq Q$. Hence, a test rejecting H_0 for 'large' values of $T_{m,n}$ will be universally consistent. However, to construct a test based on $T_{m,n}$ we need to chose a cut-off (rejection region) based on the data. The first step towards this to derive the limiting null distribution of $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$. This is discussed in Section 3.

3. Asymptotic Null Distribution

In this section we derive the limiting distribution of the vector of MMD estimates (2.10) under H_0 and, consequently, that of the proposed statistic $T_{m,n}$. In Section 3.1 we recall the definition and basic properties of multiple Weiner-Itô stochastic integrals as presented in Itô [30]. Using this framework, we derive the joint asymptotic null distribution of (2.10) in Section 3.2.

3.1. Multiple Weiner-Itô Stochastic Integrals.

Recall that $\mathcal{X}$ is a separable metric space, $\mathscr{B}(\mathcal{X})$ is the sigma-algebra generated by the open sets of $\mathcal{X}$, and P is a non-atomic probability measure on $\mathcal{X}$. We denote this probability space by $(\mathcal{X}, \mathscr{B}(\mathcal{X}), P)$.

Definition 3.1. A *Gaussian stochastic measure* on $(\mathcal{X}, \mathscr{B}(\mathcal{X}), P)$ is a collection of random variables $\{\mathcal{Z}_P(A) : A \in \mathscr{B}(\mathcal{X})\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mu)$ such that the following hold:

- $\mathcal{Z}_P(A) \sim \mathcal{N}(0, P(A))$, for all $A \in \mathscr{B}(\mathcal{X})$.
- For any finite collection of disjoint sets $A_1, \cdots, A_t \in \mathscr{B}(\mathcal{X})$, the random variables $\{\mathcal{Z}_P(A_1), \mathcal{Z}_P(A_2), \ldots, \mathcal{Z}_P(A_t)\}$ are independent and

$$\mathcal{Z}_P\left(\bigcup_{s=1}^{t} A_s\right) = \sum_{s=1}^{t} \mathcal{Z}_P(A_s).$$

For $d \geqslant 1$, denote by $L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$ the space of measurable functions $f : \mathcal{X}^d \to \mathbb{R}$ such that

$$\|f\|^2 := \int_{\mathcal{X}^d} |f(x_1, x_2, \ldots, x_d)|^2 \mathrm{d}P(x_1)\mathrm{d}P(x_2)\ldots, \mathrm{d}P(x_d) < \infty.$$

Define $\mathcal{E}_d \subseteq L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$ as the set of all elementary functions having the form

$$f(t_1, t_2, \ldots, t_d) = \sum_{1\leqslant i_1, i_2, \ldots, i_d \leqslant m} a_{i_1,i_2,\ldots,i_d} \mathbf{1}\{(t_1, t_2, \ldots, t_d) \in A_{i_1} \times \cdots \times A_{i_d}\}, \tag{3.1}$$

where $A_1, A_2, \ldots, A_m \in \mathscr{B}(\mathcal{X})$ are pairwise disjoint and $a_{i_1,i_2,\ldots,i_d}$ is zero if two indices are equal. The multiple Weiner-Itô integral for functions in $\mathcal{E}_d$ is defined as follows:

Definition 3.2. (Multiple Weiner-Itô integral for elementary functions) The d-dimensional Weiner-Itô stochastic integral, with respect to the Gaussian stochastic measure $\{Z_P(A), A \in \mathscr{B}(\mathcal{X})\}$, for the function $f \in \mathcal{E}_d$ in (3.1) is defined as

$$I_d(f) := \int_{\mathcal{X}^d} f(x_1, x_2, \ldots, x_d) \prod_{a=1}^{d} \mathrm{d}\mathcal{Z}_P(x_a) := \sum_{1 \leqslant i_1, i_2, \ldots, i_d \leqslant m} a_{i_1,i_2,\ldots,i_d} \mathcal{Z}_P(A_{i_1}) \times \cdots \times \mathcal{Z}(A_{i_d}).$$

The multiple Weiner-Itô integral for elementary functions satisfies the following two properties [30]:

- (Boundedness) For $f \in \mathcal{E}_d$, $\mathbb{E}[I_d(f)^2] \leqslant d! \|f\|^2 < \infty$.
- (Linearity) For $f, g \in \mathcal{E}_d$, $I_d(f+g) \overset{a.s.}{=} I_d(f) + I_d(g)$.

This shows that I_d is a bounded linear operator from $\mathcal{E}_d$ to $L^2(\Omega, \mathcal{F}, \mu)$, the collection of square-integrable random variables defined on $(\Omega, \mathcal{F}, \mu)$. Since $\mathcal{E}_d$ is dense in $L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$ (by [30, Theorem 2.1]), using the BLT theorem (see [50, Theorem I.7]) I_d can be uniquely extended to $L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$ by taking limits. This leads to the following definition:

Definition 3.3. (Multiple Weiner-Itô integral for general L_2-functions) The d-dimensional Weiner-Itô stochastic integral for a function $f \in L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$ is defined as the L_2 limit of the sequence $\{I_d(f_n)\}_{n \geqslant 1}$, where $\{f_n\}_{n \geqslant 1}$ is a sequence such that $f_n \in \mathcal{E}_d$ with $\lim_{n \to \infty} \|f_n - f\| = 0$. This is denoted by:

$$I_d(f) := \int_{\mathcal{X}^d} f(x_1, x_2, \ldots, x_d) \prod_{a=1}^{d} \mathrm{d}\mathcal{Z}_P(x_a). \tag{3.2}$$

As in the case of elementary functions, it can be easily checked that $I_d(f)$ satisfies the following properties:

- (Boundedness) For $f \in L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$, $\mathbb{E}[I_d(f)^2] \leqslant d! \|f\|^2 < \infty$.
- (Linearity) For $f, g \in L^2(\mathcal{X}^d, \mathscr{B}(\mathcal{X}^d), P^d)$, $I_d(f+g) \overset{a.s.}{=} I_d(f) + I_d(g)$.

It is also important to note that multiple Weiner-Itô integrals do not behave like classical (non-stochastic) integrals with respect to product measures, since by definition diagonal sets do not contribute to Itô integrals. Nevertheless, one can express the multiple Weiner-Itô integral for a product function in terms of univariate stochastic integrals using the Wick product (cf. [31, Theorem 7.26]). In the bivariate case, with 2 functions $f, g \in L^2(\mathcal{X}^2, \mathscr{B}(\mathcal{X}^2), P^2)$, this simplifies to

$$\int_{\mathcal{X}} \int_{\mathcal{X}} f(x) g(y) \mathrm{d}\mathcal{Z}_P(x) \mathrm{d}\mathcal{Z}_P(y) = \int_{\mathcal{X}} f(x) \mathrm{d}\mathcal{Z}_P(x) \int_{\mathcal{X}} g(y) \mathrm{d}\mathcal{Z}_P(y) - \int_{\mathcal{X}} f(x) g(x) \mathrm{d}x. \tag{3.3}$$

3.2. **Joint Distribution for Multiple Kernels Under H_0.** Using the framework of multiple stochastic integrals we can now describe the limiting joint distribution of the vector of MMD estimates $\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$ (recall (2.10)) under H_0. The proof is given in Section A.1.

Theorem 3.1. *Suppose $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ be a collection of r distinct kernels such that K_a satisfies Assumption 2.1 and $\mathsf{K}_a \in L^2(\mathcal{X}^2, P^2)$, for $1 \leqslant a \leqslant r$. Then under H_0, in the asymptotic regime (2.8),*

$$(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{D} G_{\mathcal{K}} := \frac{1}{\rho(1-\rho)} \Big(I_2(\mathsf{K}_1^\circ), I_2(\mathsf{K}_2^\circ), \ldots, I_2(\mathsf{K}_r^\circ) \Big)^\top, \tag{3.4}$$

where K_a° *is defined in* (2.13)*, for* $1 \leqslant a \leqslant r$*. Moreover, the characteristic function of* $G_{\mathcal{K}}$ *at* $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_r)^\top \in \mathbb{R}^r$ *is given by:*

$$\Phi(\boldsymbol{\eta}) := \mathbb{E}[e^{\iota \boldsymbol{\eta}^\top G_{\mathcal{K}}}] = \prod_{\lambda \in \Lambda(\boldsymbol{\eta})} \frac{\exp\left(-\frac{\iota\lambda}{\rho(1-\rho)}\right)}{\sqrt{1 - \frac{2\iota\lambda}{\rho(1-\rho)}}}, \tag{3.5}$$

where $\Lambda(\boldsymbol{\eta})$ *is the set of eigenvalues (with repetitions) of the Hilbert-Schmidt operator* $\mathcal{H}_{\mathcal{K},\boldsymbol{\eta}} : L^2(\mathcal{X}, P) \to L^2(\mathcal{X}, P)$ *defined as:*

$$\mathcal{H}_{\mathcal{K},\boldsymbol{\eta}}[f(x)] = \int_{\mathcal{X}} \left(\sum_{a=1}^{r} \eta_a \mathsf{K}_a^\circ(x, y) \right) f(y) \mathrm{d}P(y). \tag{3.6}$$

To prove this theorem we first apply the Cramér-Wold device and the asymptotic distribution of the univariate MMD estimate to show that linear projections of $(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$ converge to infinite weighted sums of centered χ_1^2 random variables. This representation allows us to compute the characteristic function in (3.5). Then using properties of stochastic integrals we identify (3.5) with the distribution in (3.4).

Remark 3.1 (Alternative description of the limiting distribution)**.** Note that, for $1 \leqslant a \leqslant r$, by the spectral theorem (see [51, Theorem 6.35]):

$$\mathsf{K}_a^\circ(x, y) = \sum_{s=1}^{\infty} \lambda_{s,a} \phi_{s,a}(x) \phi_{s,a}(y),$$

where $\{\lambda_{s,a}\}_{s \geqslant 1}$ and $\{\phi_{s,a}\}_{s \geqslant 1}$ are, respectively, the eigenvalues and the eigenvectors of the operator: $\mathcal{H}_{\mathsf{K}_a}[f(x)] = \int_{\mathcal{X}} \mathsf{K}_a^\circ(x, y) f(y) \mathrm{d}P(y)$. Hence, by the linearity of the stochastic integral, (3.3), and orthonormality of the eigenvectors,

$$\begin{aligned} I_2(\mathsf{K}_a^\circ) &= \sum_{s=1}^{\infty} \lambda_{s,a} \int_{\mathcal{X}} \int_{\mathcal{X}} \phi_{s,a}(x) \phi_{s,a}(y) \mathrm{d}\mathcal{Z}_P(x) \mathrm{d}\mathcal{Z}_P(y) \\ &= \sum_{s=1}^{\infty} \lambda_{s,a} \left(\left(\int_{\mathcal{X}} \phi_{s,a}(x) \mathrm{d}\mathcal{Z}_P(x) \right)^2 - \int_{\mathcal{X}} \phi_{s,a}(x)^2 \mathrm{d}x \right) \\ &\overset{D}{=} \sum_{s=1}^{\infty} \lambda_{s,a} \left(Z_{s,a}^2 - 1 \right), \end{aligned} \tag{3.7}$$

where $Z_{s,a} \overset{D}{:=} \int_{\mathcal{X}} \phi_{s,a}(x) \mathrm{d}\mathcal{Z}_P(x)$. Note that $\{Z_{s,a}\}_{s \geqslant 1, 1 \leqslant a \leqslant r}$ is a collection of Gaussian variables with

$$\mathrm{Cov}(Z_{s,a}, Z_{s',b}) = \int_{\mathcal{X}} \phi_{s,a}(x) \phi_{s',b}(x) \mathrm{d}\mathcal{Z}_P(x), \tag{3.8}$$

for $1 \leqslant a, b \leqslant r$ and $s, s' \geqslant 1$. Hence, the limiting distribution of $(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$ can be alternately expressed as (recall (3.4)):

$$G_{\mathcal{K}} \overset{D}{=} \left(\sum_{s=1}^{\infty} \lambda_{s,a} \left(Z_{s,a}^2 - 1 \right) \right)_{1 \leqslant a \leqslant r}.$$

Note that the orthonormality conditions imply that for each fixed $a \in \{1, 2, \ldots, r\}$, the collection $\{Z_{s,a}\}_{s \geqslant 1}$ is distributed as i.i.d. $\mathcal{N}(0, 1)$. This gives the well-known representation of the marginal distribution of the MMD estimate as an infinite weighted sum of independent centered

χ^2_1 variables (see [23, Theorem 12]). Jointly these infinite sums are dependent, due to the dependence among the collection $\{Z_{s,a}\}_{s\geqslant 1, 1\leqslant a\leqslant r}$ for $1 \leqslant a \neq b \leqslant r$ with covariance structure as in (3.8).

Theorem 3.1 allows us to obtain the limiting distribution of any smooth function of finitely many MMD estimates under H_0. In particular, for the MMMD statistic $T_{m,n}$ in (2.16) we have the following result:

Corollary 3.1. *Suppose* $\mathbf{\Sigma}_{H_0} := ((\sigma_{ab}))_{1\leqslant a,b\leqslant r}$ *and* $\hat{\mathbf{\Sigma}} := ((\hat{\sigma}_{ab}))_{1\leqslant a,b\leqslant r}$ *be as in* (2.12) *and* (2.14)*, respectively. Then*

$$\sigma_{ab} = \frac{2}{\rho^2(1-\rho)^2}\mathbb{E}_{X,X'\sim P}\left[\mathsf{K}^\circ_a(X,X')\mathsf{K}^\circ_b(X,X')\right], \tag{3.9}$$

where K°_a*, for* $1 \leqslant a \leqslant r$*, is as defined in* (2.13)*. Moreover, in the asymptotic regime* (2.8)*,*

$$\hat{\sigma}_{ab} \xrightarrow{a.s.} \sigma_{ab}, \tag{3.10}$$

for $1 \leqslant a, b \leqslant r$*. Furthermore, under* H_0*,*

$$(m+n)^2 T_{m,n} \xrightarrow{D} G_{\mathcal{K}}^\top \mathbf{\Sigma}_{H_0}^{-1} G_{\mathcal{K}}, \tag{3.11}$$

for $G_{\mathcal{K}}$ *as in* (3.4)*.*

The proof of Corollary 3.1 is given in Appendix A.2. It follows from Theorem 3.1 together with Slutsky's theorem and the strong law of large number for U-statistics.

4. Calibration Using Gaussian Multiplier Bootstrap

In order to apply Corollary 3.1 to obtain a valid level α tests based on $T_{m,n}$ we need to estimate the quantiles of the limiting distribution in (3.11), which depends on the (unknown) distribution P. Although the distribution in (3.11) does not have a tractable closed form, we can efficiently estimate its quantiles based on the samples $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$, using the kernel Gram matrix representation of the MMD estimate and the Gaussian multiplier bootstrap. Towards this, for each kernel K_a define its Gram matrix based on $\mathscr{X}_m$ as:

$$\hat{\mathbf{K}}_a = (\mathsf{K}_a(X_i, X_j))_{1\leqslant i,j\leqslant m},$$

and their centered versions as:

$$\hat{\mathbf{K}}^\circ_a = \boldsymbol{C}\hat{\mathbf{K}}_a\boldsymbol{C}/m = \left(\frac{\hat{\mathsf{K}}^\circ_a(X_i,X_j)}{m}\right)_{1\leqslant i,j\leqslant m}, \quad \text{where} \quad \boldsymbol{C} = \boldsymbol{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top, \tag{4.1}$$

and $\hat{\mathsf{K}}^\circ_a$ is as defined in (2.15), for $1 \leqslant a \leqslant r$. For $\hat{\rho} := \frac{m}{m+n}$, denote

$$\mathcal{E}(\mathcal{K}, \mathscr{X}_m) := \begin{pmatrix} \boldsymbol{Z}_m^\top \hat{\mathbf{K}}^\circ_1 \boldsymbol{Z}_m - \frac{1}{\hat{\rho}(1-\hat{\rho})}\operatorname{Tr}[\hat{\mathbf{K}}^\circ_1] \\ \boldsymbol{Z}_m^\top \hat{\mathbf{K}}^\circ_2 \boldsymbol{Z}_m - \frac{1}{\hat{\rho}(1-\hat{\rho})}\operatorname{Tr}[\hat{\mathbf{K}}^\circ_2] \\ \vdots \\ \boldsymbol{Z}_m^\top \hat{\mathbf{K}}^\circ_m \boldsymbol{Z}_m - \frac{1}{\hat{\rho}(1-\hat{\rho})}\operatorname{Tr}[\hat{\mathbf{K}}^\circ_m] \end{pmatrix}, \tag{4.2}$$

where $\boldsymbol{Z}_m \sim \mathcal{N}_m(\mathbf{0}, \frac{1}{\hat{\rho}(1-\hat{\rho})}\boldsymbol{I})$ independent of $\mathscr{X}_m$. In the following theorem we show that distribution of $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)$ conditional on $\mathscr{X}_m$ converges to $G_{\mathcal{K}}$ as in (3.4).

Theorem 4.1. *Suppose $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ be a collection of $r \geqslant 1$ distinct kernels such that K_a satisfies Assumption 2.1, $\mathsf{K}_a \in L^2(\mathcal{X}^2, P^2)$, and $\mathbb{E}_{X\sim P}\left[\mathsf{K}_a(X,X)^2\right] < \infty$, for $1 \leqslant a \leqslant r$. Then under H_0, in the asymptotic regime (2.8),*

$$\mathcal{E}(\mathcal{K}, \mathscr{X}_m)\big|\mathscr{X}_m \xrightarrow{D} G_{\mathcal{K}},$$

almost surely, where $G_{\mathcal{K}}$ is as defined in (3.4).

The proof of Theorem 4.1 is given in Appendix B. It shows that the asymptotic the distribution of $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)\big|\mathscr{X}_m$ is the same as that of $(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$. Since $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)\big|\mathscr{X}_m$ is completely determined by the data $\mathscr{X}_m$, we can use it approximate the quantiles of any 'nice' functions $G_{\mathcal{K}}$. To this end, define

$$\hat{T}_m := \mathcal{E}(\mathcal{K}, \mathscr{X}_m)^\top \hat{\mathbf{\Sigma}}^{-1} \mathcal{E}(\mathcal{K}, \mathscr{X}_m) \tag{4.3}$$

Now, a direct computation shows that

$$\mathrm{Var}\left[\mathcal{E}(\mathcal{K}, \mathscr{X}_m)|\mathscr{X}_m\right] = \left(\left(\mathrm{Tr}\left[\hat{\mathbf{K}}_a^\circ \hat{\mathbf{K}}_b^\circ\right]\right)\right)_{1\leqslant a,b\leqslant m}.$$

Hence, from the proof of Corollary 3.1 (specifically (3.10)) it follows that $\mathrm{Var}\left[\mathcal{E}(\mathcal{K}, \mathscr{X}_m)|\mathscr{X}_m\right] = \hat{\mathbf{\Sigma}} \xrightarrow{a.s.} \mathbf{\Sigma}_{H_0}$. This combined with Theorem 4.1 implies that under H_0,

$$\hat{T}_m \mid \mathscr{X}_m \xrightarrow{D} G_{\mathcal{K}}^\top \mathbf{\Sigma}_{H_0}^{-1} G_{\mathcal{K}}, \tag{4.4}$$

almost surely. This shows that $\hat{T}_m$ has the same limiting distribution as $(m+n)^2 T_{m,n}$ under H_0 (recall (3.11)), hence, we can use the quantiles of $\hat{T}_m$ to calibrate the statistic $T_{m,n}$. Specifically, for $\alpha \in (0,1)$ denote by $\hat{q}_{\alpha,m}$ the α-th quantile of distribution $\hat{T}_m|\mathscr{X}_m$ and consider the test function

$$\phi_{m,n} = \mathbf{1}\{(m+n)^2 T_{m,n} > \hat{q}_{1-\alpha,m}\}. \tag{4.5}$$

Corollary 3.1, (2.17), and (4.3) now implies the following result:

Corollary 4.1 (Consistency)**.** *Suppose the assumptions of Theorem 4.1 hold and $\phi_{m,n}$ be as defined above. Then $\lim_{m,n\to\infty} \mathbb{E}_{H_0}[\phi_{m,n}] = \alpha$. Moreover, for any $P \neq Q$, $\lim_{m,n\to\infty} \mathbb{E}_{H_1}[\phi_{m,n}] = 1$, that is, $\phi_{m,n}$ is universally consistent.*

The result above shows that the MMMD statistic with cut-off chosen using the multiplier bootstrap method attains the exact asymptotic level and is universally consistent. In practice, to compute $\hat{q}_{1-\alpha,m}$ we generate B replicates $\{\hat{T}_m^{(1)}, \hat{T}_m^{(2)}, \ldots, \hat{T}_m^{(B)}\}$ of $\hat{T}_m$, based on B independent copies of $\boldsymbol{Z}_m$, and choose $\hat{q}_{1-\alpha,m}$ to be the sample α-th quantile of $\{\hat{T}_m^{(1)}, \hat{T}_m^{(2)}, \ldots, \hat{T}_m^{(B)}\}$.

Remark 4.1. While implementing the test we often replace $\hat{\mathbf{\Sigma}}^{-1}$ in (2.16) and (4.3), by $(\hat{\mathbf{\Sigma}} + \lambda \boldsymbol{I}_m)^{-1}$, for some suitably chosen regularization parameter $\lambda > 0$. Although the limiting covariance matrix $\mathbf{\Sigma}_{H_0}$ is invertible (see Corollary E.1), hence, $\hat{\mathbf{\Sigma}}$ is also invertible for large sample sizes with probability 1, adding a small regularization provides numerical stability in finite samples. In fact, the conclusions in Corollary 4.1 remain valid, for any choice of $\lambda = \lambda(\mathscr{X}_m)$ converging almost surely to a deterministic constant $\lambda_0 > 0$ (see Section 7 for more details on the choice of λ in our experiments).

5. Local Asymptotic Power

In this section we derive the asymptotic power of the test based on $T_{m,n}$ against local contiguous alternatives. Throughout this section we will assume that $\mathcal{X} = \mathbb{R}^d$ and the distributions P and Q have densities f_P and f_Q with respect to the Lebesgue measure in $\mathbb{R}^d$. To quantify the notion of local alternatives, we will adopt the commonly used contamination model:

$$f_Q(\cdot) = (1-\delta) f_P(\cdot) + \delta g(\cdot), \tag{5.1}$$

where $\delta \in [0,1)$ and $g \neq f_P$ is a probability density function with respect to the Lebesgue measure in $\mathbb{R}^d$ such that the following hold:

Assumption 5.1. The support of g is contained in that of $f_P(\cdot)$ and $0 < \mathrm{Var}_{X\sim P}[\frac{g(X)}{f_P(X)}] < \infty$.

Under this assumption, contiguous local alternatives are obtained by considering local perturbations of the mixing proportion δ as follows (see [39, Chapter 12]):

$$H_0 : \delta = 0 \qquad \text{versus} \qquad H_1 : \delta = h/\sqrt{N}, \tag{5.2}$$

for some $h \neq 0$ and $N = m + n$.

Theorem 5.1. *Suppose $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ be a collection of r distinct kernels such that K_a satisfies Assumption 2.1 and $\mathsf{K}_a \in L^2(\mathcal{X}^2, P^2)$, for $1 \leqslant a \leqslant r$. Then under H_1 as in (5.2), in the asymptotic regime (2.8),*

$$(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{D} G_{\mathcal{K},h} := \begin{pmatrix} \gamma I_2(\mathsf{K}_1^\circ) + 2h\sqrt{\gamma} I_1\left(\mathsf{K}_1^\circ\left[\frac{g}{f_P}\right]\right) + h^2\mu_1 \\ \gamma I_2(\mathsf{K}_2^\circ) + 2h\sqrt{\gamma} I_1\left(\mathsf{K}_2^\circ\left[\frac{g}{f_P}\right]\right) + h^2\mu_2 \\ \vdots \\ \gamma I_2(\mathsf{K}_r^\circ) + 2h\sqrt{\gamma} I_1\left(\mathsf{K}_r^\circ\left[\frac{g}{f_P}\right]\right) + h^2\mu_r \end{pmatrix}. \tag{5.3}$$

where $\gamma = \frac{1}{\rho(1-\rho)}$, $\mathsf{K}_a^\circ[\frac{g}{f_P}](x) := \int_{\mathcal{X}} \mathsf{K}_a^\circ(x,y) g(y) \mathrm{d}y$,

$$\mu_a := \mathbb{E}\left[\mathsf{K}_a^\circ(X, X') \frac{g(X)g(X')}{f_P(X) f_P(X')}\right], \tag{5.4}$$

and K_a° is defined in (2.13), for $1 \leqslant a \leqslant r$.

The proof of Theorem 5.1 is given in Section C. The following result is an immediately consequence of the above result together with the continuous mapping theorem and Corollary 3.1.

Corollary 5.1. *Under H_1 as in (5.2), in the asymptotic regime (2.8),*

$$(m+n)^2 T_{m,n} \xrightarrow{D} G_{\mathcal{K},h}^\top \boldsymbol{\Sigma}_{H_0}^{-1} G_{\mathcal{K},h}. \tag{5.5}$$

Using Corollary 5.1 we can derive the limiting local power of the test $\phi_{m,n}$ in (4.5). Specifically, suppose $F_{\mathcal{K},h}$ denotes the CDF of $G_{\mathcal{K},h}^\top \boldsymbol{\Sigma}_{H_0}^{-1} G_{\mathcal{K},h}$ and $q_{1-\alpha}$ be the $1-\alpha$-th quantile of the distribution $G_{\mathcal{K}}^\top \boldsymbol{\Sigma}_{H_0}^{-1} G_{\mathcal{K}}$. (Note that $G_{\mathcal{K},0} = G_{\mathcal{K}}$.) Since $\hat{q}_{1-\alpha,m} | \mathscr{X}_m \xrightarrow{a.s.} q_{1-\alpha}$, (5.5) implies that the asymptotic power of $\phi_{m,n}$ under H_1 as in (5.2) is given by

$$\lim_{m,n\to\infty} \mathbb{E}_{H_1}[\phi_{m,n}] = 1 - F_{\mathcal{K},h}(q_{1-\alpha}).$$

This implies, $\phi_{m,n}$ has non-trivial asymptotic (Pitman) efficiency and is rate-optimal, in the sense that,

$$\lim_{|h|\to\infty} \lim_{m,n\to\infty} \mathbb{E}_{H_1}[\phi_{m,n}] = 1.$$

6. Distribution Under Alternative

In this section we derive the asymptotic distribution of $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$ under the alternative, that is, when $P \neq Q$.

For this we write $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$ as a two-sample U-statistics as noted in [35],

$$\mathrm{MMD}^2\left[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n\right] = \frac{1}{m(m-1)} \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant m} \sum_{1 \leqslant j_1 \neq j_2 \leqslant n} \boldsymbol{h}\left(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}\right), \tag{6.1}$$

where $\boldsymbol{h}(x, x', y, y') = \left(h_a(x, x', y, y')\right)_{1 \leqslant a \leqslant r}$ and

$$h_a(x, x', y, y') = \mathsf{K}_a(x, x') + \mathsf{K}_a(y, y') - \mathsf{K}_a(x, y') - \mathsf{K}_a(x', y), \tag{6.2}$$

for $1 \leqslant a \leqslant r$. Using the Hoeffding's decomposition we can easily derive the joint distribution of $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]$ under the alternative. In this case the asymptotic distribution will be a r-dimensional multivariate normal as in Theorem 6.1 below. To express the limiting covariance matrix we need the following definitions: For $1 \leqslant a \leqslant r$, let

$$\Delta_a^{(1)}(x) := \int_{\mathcal{X}} \mathsf{K}_a(x, x') \mathrm{d}P(x') - \int_{\mathcal{X}} \mathsf{K}_a(x, y') \mathrm{d}Q(y') \tag{6.3}$$

and

$$\Delta_a^{(2)}(y) := \int_{\mathcal{X}} \mathsf{K}_a(y, y') \mathrm{d}Q(y') - \int_{\mathcal{X}} \mathsf{K}_a(x', y) \mathrm{d}P(x'). \tag{6.4}$$

Also, denote $\boldsymbol{\Delta}^{(1)}(x) := (\Delta_a^{(1)}(x))_{1 \leqslant a \leqslant r}$ and $\boldsymbol{\Delta}^{(2)}(x) := (\Delta_a^{(2)}(x))_{1 \leqslant a \leqslant r}$. Then we have the following theorem:

Theorem 6.1. *Suppose $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ be a collection of r distinct characteristic kernels and $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_r\}$ be the unit balls of their respective RHKS. Suppose $\mathsf{K}_a \in L^2(\mathcal{X}^2, P^2) \cap L^2(\mathcal{X}^2, Q^2) \cap L^2(\mathcal{X}^2, P \times Q)$, for all $1 \leqslant a \leqslant r$. Then for $P \neq Q$,*

$$\sqrt{m+n}\left(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right] - \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]\right) \xrightarrow{D} \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Sigma}_{H_1}\right),$$

where

$$\boldsymbol{\Sigma}_{H_1} := 4\left(\rho \operatorname{Var}_{X \sim P}\left[\boldsymbol{\Delta}^{(1)}(X)\right] + (1-\rho) \operatorname{Var}_{Y \sim Q}\left[\boldsymbol{\Delta}^{(2)}(Y)\right]\right). \tag{6.5}$$

and $\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right] = \left(\mathrm{MMD}^2\left[\mathcal{F}_1, P, Q\right], \cdots, \mathrm{MMD}^2\left[\mathcal{F}_r, P, Q\right]\right)^\top$.

The proof of Theorem 6.1 is given in Appendix D. Using Theorem 6.1 we can obtain the distribution of the statistic $T_{m,n}$ (recall (2.16)) under H_1. For $\boldsymbol{z} \in \mathbb{R}^r$, define $A_{m,n}(\boldsymbol{z}) = \boldsymbol{z}^\top \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{z}$. Then by a Taylor series expansion, Corollary 3.1, and Theorem 6.1,

$$\begin{aligned}
&T_{m,n} - \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right] \\
&= A_{m,n}(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right]) - A_{m,n}(\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]) \\
&= 2\left(\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n\right] - \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right] + O_P(1/N),
\end{aligned}$$

where $N := m + n$. Hence, using Theorem 6.1 and $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}_{H_0}$ (by Corollary 3.1),

$$\sqrt{m+n}\left(T_{m,n} - \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]\right) \xrightarrow{D} \mathcal{N}(0, \sigma_{H_1}^2),$$

where $\sigma_{H_1}^2 := 4\mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]^\top \boldsymbol{\Sigma}_{H_0}^{-1} \boldsymbol{\Sigma}_{H_1} \boldsymbol{\Sigma}_{H_0}^{-1} \mathrm{MMD}^2\left[\mathcal{F}, P, Q\right]$, with $\boldsymbol{\Sigma}_{H_1}$ as defined in (6.5).

7. Numerical Experiments

In this section, we study the finite-sample performance of the proposed MMMD test, both in terms of Type-I error control and power, across a range of simulation settings. Specifically, we will compare the MMMD test with the single kernel MMD test [22] and the graph-based Friedman Rafsky (FR) test [19]. Throughout we set the significance level $\alpha = 0.05$.

For single kernel tests we use the Gaussian and Laplace kernels:

$$\mathsf{K}_{\mathrm{GAUSS}}(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}} \quad \text{and} \quad \mathsf{K}_{\mathrm{LAP}}(x,y) = e^{-\frac{\|x-y\|}{\sigma}},$$

with the bandwidth σ is chosen using the median heuristic

$$\sigma^2 := \lambda^2_{\mathrm{med}} = \mathrm{median}\left\{\|Z_i - Z_j\|^2 : 1 \leqslant i < j \leqslant n\right\},$$

where $\mathscr{X}_m \cup \mathscr{Y}_n = \{Z_1, Z_2, \ldots, Z_N\}$ is the pooled sample and $\|\cdot\|$ denotes the Euclidean norm. We will refer to these tests as `Gauss MMD` and `LAP MMD`, respectively.

For the MMMMD statistic we will use multiple Gaussian kernels, multiple Laplace kernels, or combination of Gaussian and Laplace kernels, with different bandwidths chosen follows:

- `Gauss MMMD`: This is the MMMD statistic with 5 Gaussian kernels with bandwidths

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (\tfrac{1}{2}, \tfrac{1}{\sqrt{2}}, 1, \sqrt{2}, 2)\lambda_{\mathrm{med}}. \tag{7.1}$$

- `LAP MMMD`: This is the MMMD statistic with 5 Laplace kernels with bandwidths

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (\tfrac{1}{2}, \tfrac{1}{\sqrt{2}}, 1, \sqrt{2}, 2)\lambda_{\mathrm{med}}. \tag{7.2}$$

- `Mixed MMMD`: This is the MMMD statistic with 3 Gaussian kernels and 3 Laplace kernels with same set of bandwidths

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3) = (\tfrac{1}{\sqrt{2}}, 1, \sqrt{2})\lambda_{\mathrm{med}}. \tag{7.3}$$

In our implementation we choose the regularity parameter λ (recall Remark 4.1) as: $\lambda = 10^{-5} \times \min_{1 \leqslant a \leqslant r} \hat{\sigma}_{aa}$, for $\hat{\sigma}_{aa} > 0$ as in (2.14). Since λ converges to $10^{-5} \times \min_{1 \leqslant a \leqslant r} \sigma_{aa}$ almost surely (recall Corollary 3.1), the results in Corollary 4.1 remain valid. The cutoffs of the tests are chosen based on the multiplier bootstrap as in (4.5) using $B = 500$ resamples.

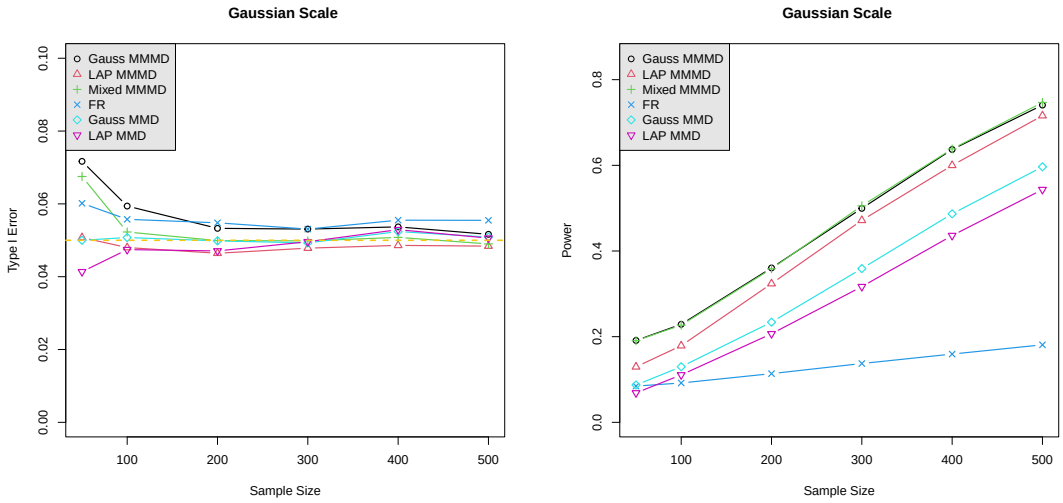
Finally, for the Friedman Rafsky (FR) test we use the implementation in the `R` package `gTests`, with the 5-MST (minimum spanning tree), which is the recommended practical choice in [11].

7.1. Dependence on Sample Size.

In this section we illustrate how the different tests performing as the sample size varies, with dimension held fixed. Toward this, we fix $d = 2$ and consider

$$P = \mathcal{N}_2(\mathbf{0}, \boldsymbol{I}_2) \text{ and } Q = \mathcal{N}_2(\mathbf{0}, 1.25 \cdot \boldsymbol{I}_2),$$

and vary the sample sizes over $m = n \in \{50, 100, 200, 300, 400, 500\}$. Figure 1 shows the empirical Type I-error and power of the aforementioned tests. Figure 1(a) shows that all the tests have good Type I error control. Figure 1(b) shows that the multiple kernel MMMD tests have better power than the single kernel MMD tests, with the `Gauss MMMD` and the `Mixed MMMD` tests performing the best.



(a) (b)

FIGURE 1. (a) Empirical type-I errors and (b) powers of the different tests in dimension $d = 2$ with varying sample size.

7.2. **Dependence on Dimension.** In this section we study the performance of the different tests as dimension varies in the following 4 settings. We fix sample sizes $m = n = 100$, vary dimension over $d \in \{5, 10, 25, 50, 75, 100, 150\}$, and compute the empirical power by averaging over 500 iterations.

(S1) *Gaussian location-scale*: Here, we consider $P = \mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma}_0)$ and $Q = \mathcal{N}_d(0.1\mathbf{1}, 1.15\mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0 = ((0.5^{|i-j|}))_{1 \leqslant i,j \leqslant d}$ (see Figure 2(a)).

(S2) *t-distribution scale*: Here, we consider $P = t_{10}(\mathbf{0}, \mathbf{\Sigma}_0)$ and $Q = t_{10}(\mathbf{0}, 1.22\mathbf{\Sigma}_0)$, where t_{10} is the t-distribution with 10 degrees of freedom and $\mathbf{\Sigma}_0$ is as above (see Figure 2(b)).

(S3) *Gaussian and t-distribution mixture*: Here, we consider

$$P = \tfrac{1}{2}\mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma}_0) + \tfrac{1}{2}t_{10}(\mathbf{0}, \mathbf{\Sigma}_0) \text{ and } Q = \tfrac{1}{2}\mathcal{N}_d(\mathbf{0}, 1.22\mathbf{\Sigma}_0) + \tfrac{1}{2}t_{10}(\mathbf{0}, 1.22\mathbf{\Sigma}_0),$$

where $\mathbf{\Sigma}_0$ is as above (see Figure 3(a)).

(S4) *Gaussian and Laplace mixture*: Here, we consider

$$P = \tfrac{1}{2}\mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma}_1) + \tfrac{1}{2}t_{10}(\mathbf{0}, \mathbf{\Sigma}_1) \text{ and } Q = \tfrac{1}{2}\mathcal{N}_d(\mathbf{0}, 1.3\mathbf{\Sigma}_1) + \tfrac{1}{2}t_{10}(\mathbf{0}, 1.3\mathbf{\Sigma}_1),$$

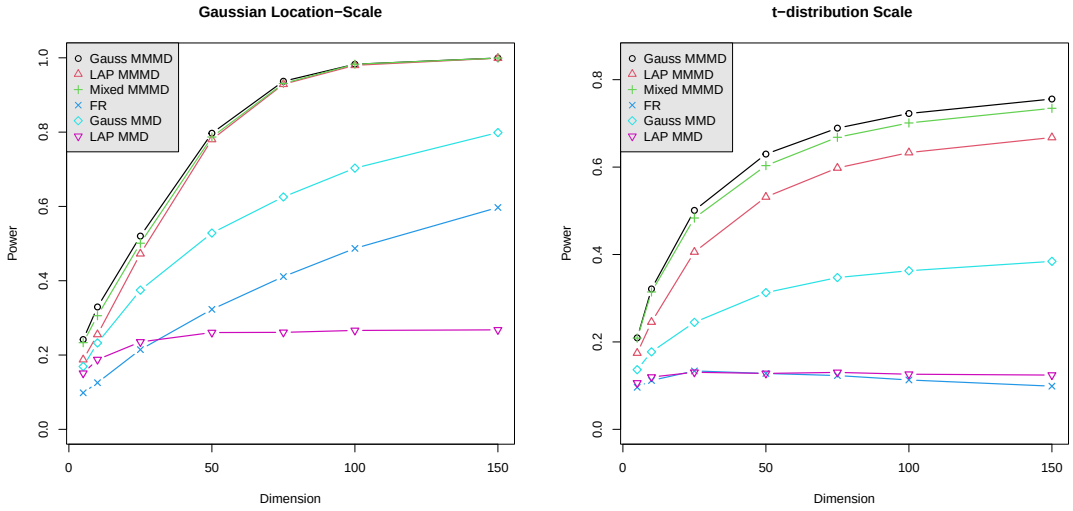
where $\mathbf{\Sigma}_1 = ((0.7^{|i-j|}))_{1 \leqslant i,j \leqslant d}$ (see Figure 3(b)).

The plots show that the multiple kernel MMMD tests have significantly more power than the single kernel MMD tests and the FR test in all the 4 settings. Overall the `Gauss MMMD` and the `Mixed MMMD` tests perform the best, closely followed by the `Lap MMMD`. This also shows the advantage of aggregating kernels across a range of dimensions, from low dimensions to dimensions that are comparable and even larger than the sample size.

7.3. **Mixture Alternatives.** In this section we evaluate the performance of the tests for mixture alternatives by varying the mixing proportion. To this end, suppose $\mathbf{\Sigma}_0 = ((0.5^{|i-j|}))_{1 \leqslant i,j \leqslant d}$ and consider

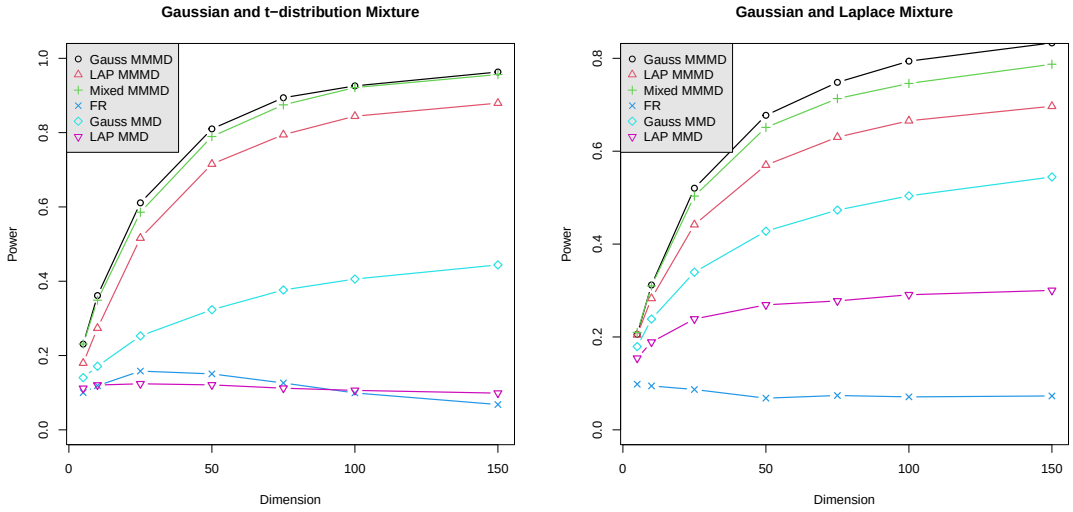
$$P = \varepsilon\mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma}_0) + (1-\varepsilon)t_{10}(\mathbf{0}, \mathbf{\Sigma}_0) \text{ and } Q = \varepsilon\mathcal{N}_d(\mathbf{0}, 1.25\mathbf{\Sigma}_0) + (1-\varepsilon)t_{10}(\mathbf{0}, 1.25\mathbf{\Sigma}_0).$$

Figure 4 shows the empirical power (averaged over 500 iterations) of the different tests as ε varies over $[0, 1]$, with sample sizes $m = n = 100$ and dimension $d = 30$ (Figure 4(a)) and



(a) (b)

FIGURE 2. Empirical powers of the different tests as the dimension varies in (a) setting **(S1)** and (b) setting **(S2)**.



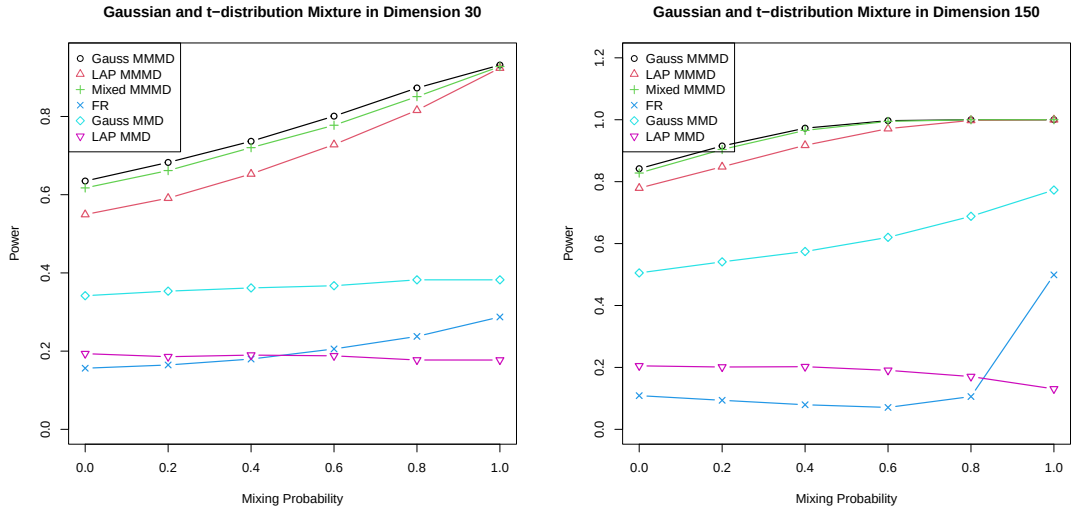
(a) (b)

FIGURE 3. Empirical powers of the different tests as the dimension varies in (a) setting **(S3)** and (b) setting **(S4)**.

$d = 150$ (Figure 4 (b)). In both cases, the MMMD tests outperform the single kernel tests and the FR test, again illustrating the versatility of the aggregated tests.

7.4. **Local Alternatives.** Recall that in Section 5 we derived the asymptotic local power of the MMMD statistic. Here we validate this in the following simulation setting:

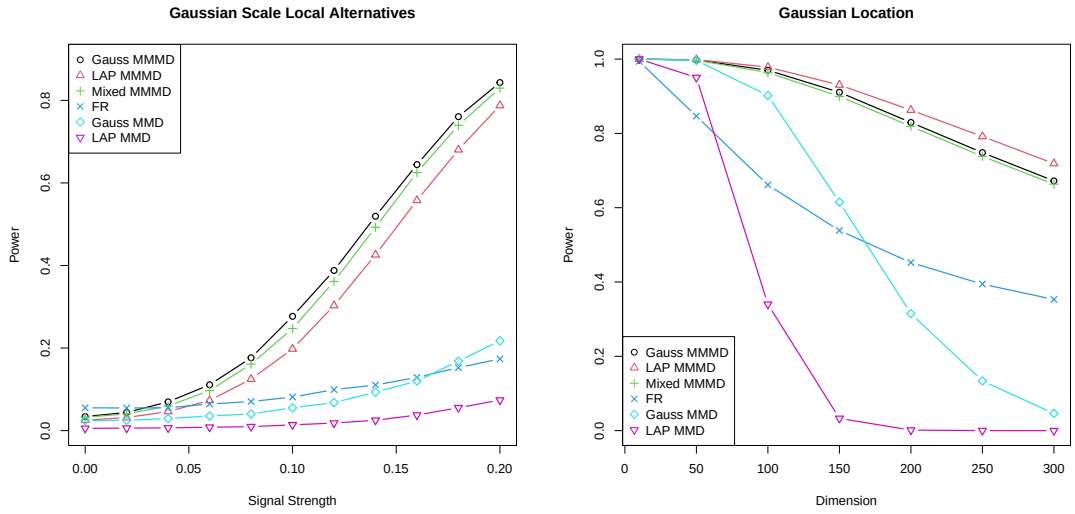
$$P = \mathcal{N}_d(\mathbf{0}, \boldsymbol{I}_d) \text{ and } Q = \mathcal{N}_d\left(\mathbf{0}, (1 + \tfrac{h}{\sqrt{N}})\boldsymbol{I}_d\right), \tag{7.4}$$



(a) (b)

FIGURE 4. Empirical powers of the different tests for mixture alternatives as a function of the mixing proportion for dimension (a) $d = 30$ and (b) $d = 150$.

where $N = m + n$. Figure 5(a) shows the empirical power (averaged over 500 iterations) of the different tests, for dimension $d = 20$, sample sizes $m = n = 100$, as the signal strength varies over [0,3]. The plots show that the MMMD methods have significantly better local power than the other tests, illustrating the attractive efficiency property of our aggregation method.



(a) (b)

FIGURE 5. Empirical powers of the different tests for (a) local alternatives as in (7.4) and (b) high-dimensional alternatives as in (7.5).

7.5. **High-Dimensional Alternatives.** To fairly evaluate the performance of kernel tests in high-dimensions we consider, as suggested by Ramdas et al. [48], pairs of distributions for which

the Kullback-Leibler (KL) divergence remain constant, as the dimension increases. Specifically, we consider following

$$P = \mathcal{N}_d\left(\mathbf{0}, \boldsymbol{I}_d\right) \text{ and } Q = \mathcal{N}_d\left((1.25/\sqrt{d})\mathbf{1}, \boldsymbol{I}_d\right). \tag{7.5}$$

It is easy to check that the KL divergence between P and Q is $\frac{1.25^2}{2}$, which does not change with d. In this case for the `Gauss MMMD`/`LAP MMMD` tests we use 5 different Gaussian/Laplace kernels with respective bandwidths $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (\frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2})\lambda_{\text{med}}$. Also, for the `Mixed MMMD` test is implemented with 4 Gaussian kernels and 4 Laplace kernels with same set of bandwidths: $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (\frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2})\lambda_{\text{med}}$. Figure 5(b) shows the empirical powers (averaged over 500 iterations) of the different tests as a function of dimension with sample sizes $m = n = 100$. As expected, the power of all the tests decreases as dimension increases. However, for the multiple kernel tests the power decrement is much slower and overall they perform significantly better than the single kernel tests.

8. Real Data Applications

In this section we apply our method to compare images of digits in the noisy MNIST dataset. Specifically, consider two noisy versions of the MNIST dataset: (1) MNIST with additive Gaussian noise (Section 8.1), and (2) MNIST with reduced contrast and additive noise, where, in addition to the Gaussian noise the contrast of the images is reduced (Section 8.2). As in the previous section, we implement the single kernel `Gauss MMD` and `LAP MMD` tests with the median bandwidth, the multiple kernel `Gauss MMMD`, `LAP MMMD`, and `Mixed MMMD` tests with bandwidths as in (7.1), (7.2), (7.3), respectively, and the FR test using the R package `gTests`.

8.1. **MNIST with Additive Gaussian Noise.** In this section we illustrate the performance of the proposed test in detecting different sets of digits when i.i.d. Gaussian noise with standard deviation σ is added to each pixel. Figure 6 shows how such noisy data looks for (a) $\sigma = 0$ (which is the clean data with no noise), (b) $\sigma = 0.6$ and (c) $\sigma = 1$.

(a) Clean data (b) Noisy data with $\sigma = 0.6$ (c) Noisy data with $\sigma = 1$

Figure 6. MNIST data with additive Gaussian noise.

To evaluate the proposed method we consider the following sets of digits:

$$P = \{1, 2, 3\} \text{ and } Q = \{1, 2, 8\},$$

and vary the standard error $\sigma \in (0, 0.2, 0.4, 0.6, 0.8, 1)$. For each σ we draw 100 samples with replacements from the two sets and check if the tests successfully reject H_0 at level $\alpha = 0.5$. We repeat this experiment 500 times to estimate the power. Figure 7 shows performance of the

above mentioned tests, where we plot the power over the index of pair of sets of digits. This shows that for the clean data and for small noise levels, the singe kernel `Gauss MMD` performs comparably with the MMMD tests. However, for larger noise levels the MMMD tests perform much than the single kernel tests. The FR test also perform well in this case across the range of the noise level.

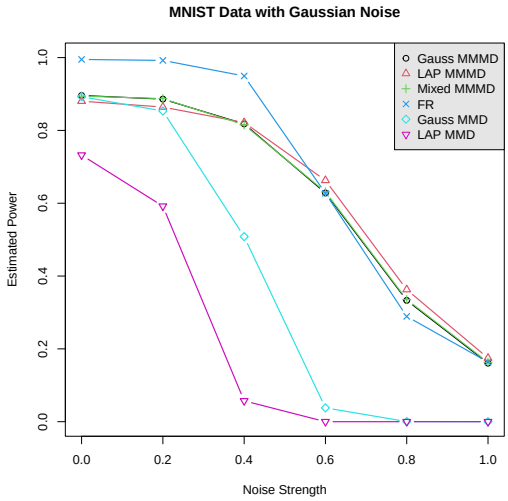


FIGURE 7. Estimated powers of the tests on noisy MNIST dataset with increasing noise strength.

8.2. **MNIST with reduced contrast and additive Gaussian noise.** In this section we illustrate the performance of the different tests on the noisy version of the MNIST dataset considered in [6]. (The dataset is publicly available at https://csc.lsu.edu/~saikat/n-mnist/) Here, in addition to additive Gaussian noise the contrast of the images is also reduced. Specifically, the contrast range is scaled down to half and an additive Gaussian noise is introduced with signal-to-noise ratio of 12. This emulates background clutter along with significant change in lighting conditions (see Figure 8 for an example of such a noisy image).

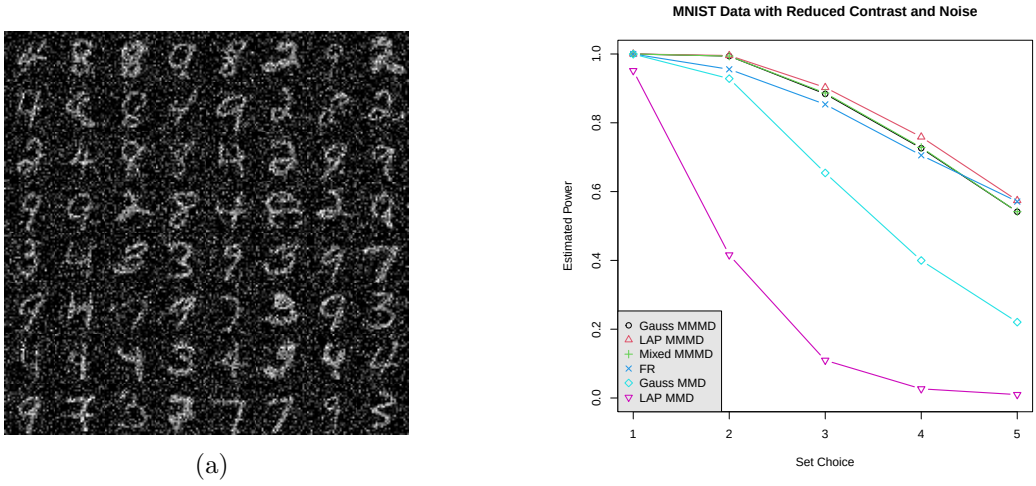
We evaluate the performance of the different test for the following 5 pairs of sets of digits:

(1) $P = \{2, 4, 8, 9\}$ and $Q = \{3, 4, 7, 9\}$,
(2) $P = \{1, 2, 4, 8, 9\}$ and $Q = \{1, 3, 4, 7, 9\}$,
(3) $P = \{0, 1, 2, 4, 8, 9\}$ and $Q = \{0, 1, 3, 4, 7, 9\}$,
(4) $P = \{0, 1, 2, 4, 5, 8, 9\}$ and $Q = \{0, 1, 3, 4, 5, 7, 9\}$,
(5) $P = \{0, 1, 2, 4, 5, 6, 8, 9\}$ and $Q = \{0, 1, 3, 4, 5, 6, 7, 9\}$.

For each of the 5 cases above, we draw 150 samples with replacements from the two sets and check if the tests successfully reject H_0 at level $\alpha = 0.5$. We repeat this experiment 500 times to estimate the power. Figure 8 shows the power of the different for the above 5 sets. In this case, the multiple kernel tests and the FR test overall has the highest power across the 5 sets, followed by the `Gauss MMD` and the `Lap MMD`.

9. Broader Scope

The idea of using multiple kernels/bandwidths has recently emerged as a popular alternative to selecting a single bandwidth, for developing adaptive kernel two-sample tests that do not require data-splitting. In this direction, Kübler et al. [37] proposed a method which does not



(a) (b)

FIGURE 8. (a) MNIST dataset with reduced contrast and additive white gaussian noise and (b) estimated power.

require data splitting using the framework of post-selection inference. However, this method requires asymptotic normality of the test statistic under H_0, hence, is restricted to the linear-time MMD estimate [23, Section 6], which leads to loss in power when compared to the more commonly used quadratic-time estimate (2.6). Fromont et al. [20, 21] and, more recently, Schrab et al. [55] introduced another non-asymptotic aggregated test, hereafter referred to as MMDAgg, that is adaptive minimax (up to an iterated logarithmic term) over Sobolev balls.

Our aggregation strategy, leads to a test that can be efficiently implemented, enjoy improved empirical power over single kernel tests for a range of alternatives, and scales well in high dimensions. Moreover, instead of minimax optimality, our focus is on establishing the asymptotic properties of the aggregated test. Towards this, we derive the joint distribution of the MMD estimates (under both local and fixed alternatives) and, consequently, establish the statistical (Pitman) efficiency of the proposed test. In fact, our theoretical results apply to general aggregation schemes using which we can obtain the asymptotic efficiency of the aforementioned MMDAgg test (see Section 9.1). Numerical results comparing the empirical power of the MMMD test with the MMDAgg test are reported in Section 9.2. Interestingly, MMMD has better power than MMDAgg for a range of alternatives, which include perturbed uniform distributions in the Sobolev class, as well as natural mixture and local alternatives. This showcases both the practical relevance of the Mahalanobis aggregation strategy and the broader scope of our asymptotic results.

9.1. **Local Asymptotic Power of the MMDAgg Test.** In this section, we propose an asymptotic implementation of the MMDAgg test and sketch a heuristic argument that derives its limiting local power in the contamination model (5.2). The argument can be made rigorous by using tools from empirical process theory, however, since the purpose of this section is more illustrative than technical, we have not pursued this direction.

To describe the asymptotic version of the MMDAgg test suppose $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ is a finite collection of kernels and $\mathcal{W} := \{w_1, w_2, \ldots, w_r\}$ is an associated collection of positive weights such that $\sum_{s=1}^{r} w_s \leqslant 1$. Moreover, for $\alpha \in (0,1)$ and $1 \leqslant s \leqslant r$, let $\hat{q}_{1-\alpha,s,m}$ be the α-th

quantile of the distribution

$$\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) := \boldsymbol{Z}_m^\top \hat{\mathsf{K}}_s^\circ \boldsymbol{Z}_m - \frac{1}{\hat{\rho}(1-\hat{\rho})} \operatorname{Tr}[\hat{\mathsf{K}}_s^\circ].$$

where $\hat{\mathsf{K}}_s^\circ$ is as defined in (4.1), for $1 \leqslant a \leqslant r$, and $\boldsymbol{Z}_m \sim \mathcal{N}_m(\boldsymbol{0}, \frac{1}{\hat{\rho}(1-\hat{\rho})}\boldsymbol{I})$ is independent of $\mathscr{X}_m$. The idea of the MMDAgg test is to reject H_0 if any one of the individual (single-kernel) test based on the kernels in $\mathcal{K}$ rejects H_0 for a specially chosen cut-off (see [55, Section 3.5] for details). Here, we consider an alternative implementation of MMDAgg test based on the Gaussian multiplier bootstrap discussed in Section 4. To this end, define

$$u^*_{\alpha,m} := \arg\max \left\{ u \in (0, L) : \mathbb{P}\left(\max_{1 \leqslant s \leqslant r} \{\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) - \hat{q}_{1-uw_s,s,m}\} > 0 | \mathscr{X}_m \right) \leqslant \alpha \right\},$$

where $L := \min_{1 \leqslant s \leqslant r} w_s^{-1}$. (Note that the probability in the RHS above is over the randomness of $\boldsymbol{Z}_m$ (conditional on $\mathscr{X}_m$), hence, $u^*_{\alpha,m}$ can be computed from the data by a grid search over $u \in (0, L)$.) The MMDAgg test would then reject H_0 if

$$\phi^{\mathrm{MMDAgg}}_{m,n,\alpha} := \mathbf{1}\left\{ \max_{1 \leqslant s \leqslant r} \left\{ \mathrm{MMD}^2\left[\mathsf{K}_s, \mathscr{X}_m, \mathscr{Y}_n\right] - \hat{q}_{1-w_s u^*_{\alpha},s,m} \right\} > 0 \right\}. \tag{9.1}$$

To describe the asymptotic properties of this test, let $q_{\alpha,s}$ be the α-th quantile of the distribution $\frac{1}{\rho(1-\rho)} I_2(\mathsf{K}_s^\circ)$, for $1 \leqslant s \leqslant r$. Then for each fixed $u \in (0, L)$, by Theorem 3.1, Slutsky's theorem, and the continuous mapping theorem, as $m \to \infty$,

$$\max_{1 \leqslant s \leqslant r} \{\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) - \hat{q}_{1-uw_s,s,m}\} \xrightarrow{D} \max_{1 \leqslant s \leqslant r} \left\{ \frac{1}{\rho(1-\rho)} I_2(\mathsf{K}_s^\circ) - q_{1-uw_s,s} \right\}.$$

since $\hat{q}_{1-uw_s,s,m} | \mathscr{X}_m \xrightarrow{a.s.} q_{1-uw_s,s}$. Therefore, for each fixed $u \in (0, L)$, as $m \to \infty$,

$$\mathbb{P}\left(\max_{1 \leqslant s \leqslant r} \{\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) - \hat{q}_{1-uw_s,s,m}\} > 0 | \mathscr{X}_m \right) \to \mathbb{P}\left(\max_{1 \leqslant s \leqslant r} \left\{ \frac{1}{\rho(1-\rho)} I_2(\mathsf{K}_s^\circ) - q_{1-uw_s,s} \right\} > 0 \right).$$

Now, since the convergence of the quantiles is uniform, we expect the following to hold as $m \to \infty$:

$$u^*_{\alpha,m} \xrightarrow{a.s.} u^*_\alpha \quad \text{and} \quad \hat{q}_{1-w_s u^*_{\alpha,m},s,m} | \mathscr{X}_m \xrightarrow{a.s.} q_{1-w_s u^*_\alpha,s},$$

where

$$u^*_\alpha := \arg\max \left\{ u \in (0, L) : \mathbb{P}\left(\max_{1 \leqslant s \leqslant r} \left\{ \frac{1}{\rho(1-\rho)} I_2(\mathsf{K}_s^\circ) - q_{1-uw_s,s} \right\} > 0 \right) \leqslant \alpha \right\}.$$

Hence, under H_1 as in (5.2), by Theorem 5.1, Slutsky's theorem, and the continuous mapping theorem,

$$\max_{1 \leqslant s \leqslant r} \left\{ \mathrm{MMD}^2\left[\mathsf{K}_s, \mathscr{X}_m, \mathscr{Y}_n\right] - \hat{q}_{1-w_s u^*_{\alpha},s,m} \right\} \xrightarrow{D} \max_{1 \leqslant s \leqslant r} \left\{ G_{\mathsf{K}_s,h} - q_{1-w_s u^*_{\alpha},s} \right\},$$

where

$$G_{\mathsf{K}_s,h} := \gamma I_2(\mathsf{K}_s^\circ) + 2h\sqrt{\gamma} I_1\left(\mathsf{K}_s^\circ \left[\frac{g}{f_P} \right] \right) + h^2 \mu_s,$$

and μ_s is as defined in (5.4). Therefore, the limiting power of the test (9.1) is given by

$$\begin{aligned} \lim_{m,n \to \infty} \mathbb{E}_{H_1}[\phi^{\mathrm{MMDAgg}}_{m,n,\alpha}] &= \mathbb{P}\left(\max_{1 \leqslant s \leqslant r} \left\{ G_{\mathsf{K}_s,h} - q_{1-w_s u^*_{\alpha},s} \right\} > 0 \right) \\ &= 1 - \mathbb{P}\left(G_{\mathsf{K}_s,h} \leqslant q_{1-w_s u^*_{\alpha},s}, \text{ for all } 1 \leqslant s \leqslant r \right) \\ &= 1 - \boldsymbol{F}_{\mathcal{K},h}(q_{1-w_1 u^*_{\alpha},1}, \ldots, q_{1-w_r u^*_{\alpha},r}), \end{aligned}$$

where $\boldsymbol{F}_{\mathcal{K},h}$ is the cumulative distribution function of the vector $(G_{\mathsf{K}_1,h}, G_{\mathsf{K}_2,h}, \ldots, G_{\mathsf{K}_r,h})^\top$.

9.2. **Empirical Comparison.** In this section we compare the MMMD test with the MMDAgg test as implemented in [55].

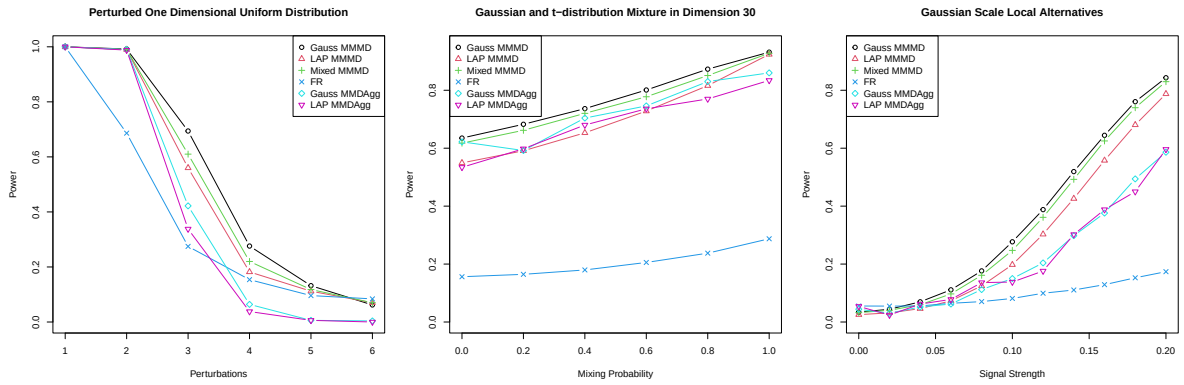
9.2.1. *Perturbed* 1*-dimensional Uniform Distribution.* First we consider a perturbed version of uniform distribution on $[0,1]$ as considered [55]. The perturbed density at $x \in \mathbb{R}$ is given by:

$$f_{\boldsymbol{\theta}}(x) = \mathbf{1}\{x \in [0,1]\} + \frac{c_1}{P} \sum_{v \in \{1,2,\ldots,P\}} \theta_v G(Px - v)$$

where $c_1 = 2.7$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_P) \in \{-1,1\}^P$, P is the number of perturbations, and

$$G(t) := \exp\left(-\frac{1}{1-(4t+3)^2}\right)\mathbf{1}\left\{t \in (-1,-\tfrac{1}{2})\right\} - \exp\left(-\frac{1}{1-(4t+1)^2}\right)\mathbf{1}\left\{t \in (-\tfrac{1}{2},0)\right\}.$$

It is known that for P large enough the difference between the uniform density and the perturbed uniform density lies in the Sobolev ball [40]. Figure 9(a) shows the empirical powers of the `MMDAgg` test with Gaussian and Laplace kernels, with bandwidths chosen according to the increasing weight strategy as in [55]; the empirical powers of the `Gauss MMMD`, `LAP MMMD`, and `Mixed MMMD` with bandwidth chosen as in (7.1), (7.2) and (7.3), respectively; and the empirical power of the FR test. The sample sizes are set to $m = n = 500$, the perturbations range over $P = 1,2,3,4,5,6$, and the power is computed over 500 repetitions, with a new value of $\boldsymbol{\theta} \in \{-1,1\}^P$ sampled uniformly in each iteration. The plot shows that the MMMD tests have better finite-sample power than the MMAgg tests, particularly for larger perturbations.



(a) (b) (c)

FIGURE 9. Empirical powers of the MMMD and MMDAgg tests for (a) the perturbed uniform distribution, (b) mixture alternatives, and (c) local alternatives.

9.2.2. *Mixture and Local Alternatives.* Next, we compare the empirical power (by repeating the experiment 500 times) of the MMMD tests with the MMDAgg tests (based on Gaussian and Laplace kernels) for the mixtures alternative in $d = 30$ as in Section 7.3 and the local alternative in $d = 20$ as in Section 7.4. For the MMMD tests we use bandwidths as in (7.1), (7.2), and (7.3), while for the MMDAgg tests we consider the increasing weight strategy with collection of bandwidths $\Lambda(-2,2)$ as defined in [55, Section 5.3]. For the mixture alternative the MMMD tests perform slightly better than the MMDAgg tests (see Figure 9(b)), while for the local alternative the MMMD tests show significant improvement over the MMDAgg tests (see Figure 9(c)).

References

[1] T. W. Anderson. On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159, 1962.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[3] B. Aslan and G. Zech. New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation*, 75(2):109–119, 2005.

[4] B. Banerjee and A. K. Ghosh. On high dimensional behaviour of some two-sample tests based on ball divergence. *arXiv preprint arXiv:2212.08566*, 2022.

[5] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.

[6] S. Basu, M. Karki, S. Ganguly, R. DiBiano, S. Mukhopadhyay, S. Gayaka, R. Kannan, and R. Nemani. Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Processing Letters*, 45(3):855–867, 2017.

[7] B. B. Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):575–602, 2019.

[8] B. B. Bhattacharya, P. Diaconis, and S. Mukherjee. Universal limit theorems in graph coloring problems with connections to extremal combinatorics. *The Annals of Applied Probability*, 27(1): 337–394, 2017.

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

[10] M. Biswas, M. Mukhopadhyay, and A. K. Ghosh. A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101(4):913–926, 2014.

[11] H. Chen and J. H. Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409, 2017.

[12] M. Chikkagoudar and B. V. Bhat. Limiting distribution of two-sample degenerate U-statistic under contiguous alternatives and applications. *Journal of Applied Statistical Science*, 22(1/2):127, 2014.

[13] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015.

[14] N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, pages 1–16, 2021.

[15] N. Deb, B. B. Bhattacharya, and B. Sen. Efficiency lower bounds for distribution-free Hotelling-type two-sample tests based on optimal transport. *arXiv:2104.01986*, 2021.

[16] N. Dunford and J. T. Schwartz. *Linear operators. Part II.* Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988.

[17] R. Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

[18] H. Finner. A generalization of Holder's inequality and some probability inequalities. *The Annals of Probability*, pages 1893–1901, 1992.

[19] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

[20] M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Conference on Learning Theory*, pages 23–1. JMLR Workshop and Conference Proceedings, 2012.

[21] M. Fromont, B. Laurent, and P. Reynaud-Bouret. The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3): 1431–1461, 2013.

[22] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS*, volume 23, pages 673–681, 2009.

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

[24] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213. Citeseer, 2012.

[25] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[26] R. Heller, S. T. Jensen, P. R. Rosenbaum, and D. S. Small. Sensitivity analysis for the cross-match test, with applications in genomics. *Journal of the American Statistical Association*, 105(491):1005–1013, 2010.

[27] R. Heller, P. R. Rosenbaum, and D. S. Small. Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, 64(4):299–309, 2010.

[28] N. Henze. On the number of random points with nearest neighbour of the same type and a multivariate two-sample test. *Metrika*, 31:259–273, 1984.

[29] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20(1):37 – 39, 1953.

[30] K. Itô. Multiple wiener integral. *Journal of the Mathematical Society of Japan*, 3(1):157–169, 1951.

[31] S. Janson. *Gaussian Hilbert spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997.

[32] S. Janson and K. Nowicki. The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probab. Theory Related Fields*, 90(3):341–375, 1991.

[33] I. Kim, S. Balakrishnan, and L. Wasserman. Robust multivariate nonparametric tests via projection averaging. *The Annals of Statistics*, 48(6):3417–3441, 2020.

[34] I. Kim, A. Ramdas, A. Singh, and L. Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434, 2021.

[35] I. Kim, S. Balakrishnan, and L. Wasserman. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251, 2022.

[36] V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[37] J. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet. Learning kernel tests without data splitting. *Advances in Neural Information Processing Systems*, 33:6245–6255, 2020.

[38] P. D. Lax. *Functional Analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2002.

[39] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[40] T. Li and M. Yuan. On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*, 2019.

[41] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pages 6316–6326. PMLR, 2020.

[42] R. Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1): 405–414, 1990.

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

[44] D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[45] E. Lubetzky and Y. Zhao. On replica symmetry of large deviations in random graphs. *Random Structures & Algorithms*, 47(1):109–146, 2015.

[46] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60, 1947.

[47] W. Pan, Y. Tian, X. Wang, and H. Zhang. Ball divergence: nonparametric two sample test. *The Annals of Statistics*, 46(3):1109–1137, 2018.

[48] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[49] A. Ramdas, N. García Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.
[50] M. Reed and B. Simon. *Methods of Modern Mathematical Physics. vol. 1. Functional Analysis*. Academic New York, 1980.
[51] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004.
[52] P. R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4): 515–530, 2005.
[53] S. Sarkar, R. Biswas, and A. K. Ghosh. On some graph-based two-sample tests for high dimension, low sample size data. *Machine Learning*, 109(2):279–306, 2020.
[54] M. F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.
[55] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD aggregated two-sample test. *arXiv:2110.15073*, 2021.
[56] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.
[57] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1980.
[58] S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel two-sample test. *arXiv:2211.14908*, 2022.
[59] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19:279–281, 1948.
[60] H. Song and H. Chen. Generalized kernel two-sample tests. *arXiv:2011.06127*, 2020.
[61] G. J. Székely. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18, 2003.
[62] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10): 1249–1272, 2004.
[63] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.
[64] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.
[65] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11:147–162, 1940.
[66] L. Weiss. Two-sample tests for multivariate distributions. *The Annals of Mathematical Statistics*, pages 159–164, 1960.
[67] F. Wilcoxon. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3: 119–122, 1947.
[68] J.-T. Zhang, J. Guo, and B. Zhou. Testing equality of several distributions in separable metric spaces: A maximum mean discrepancy based approach. *Journal of Econometrics*, 2022.

Appendix A. Proof of Theorem 3.1 and Corollary 3.1

To begin with note that the definition of MMD^2 in (2.5) can be extended to any measurable and symmetric function $\mathsf{H} \in L^2(\mathcal{X}^2, P^2)$ (not necessarily positive definite) in a natural way as follows:

$$\mathrm{MMD}^2\left[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n\right] = \mathcal{W}_{\mathscr{X}_m} + \mathcal{W}_{\mathscr{Y}_n} - 2\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n}, \tag{A.1}$$

where $\mathcal{W}_{\mathscr{X}_m}$, $\mathcal{W}_{\mathscr{Y}_n}$, and $\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n}$ are as defined in (2.6) and (2.7), respectively, with K replaced by H. The main ingredient in the proof of Theorem 3.1 is the following result:

Proposition A.1 ([23, Theorem 5]). *For any measurable and symmetric function* $\mathsf{H} \in L^2(\mathcal{X}^2, P^2)$, *in the asymptotic regime* (2.8),

$$(m+n)\mathrm{MMD}^2\left[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n\right] := \frac{1}{\rho(1-\rho)} \sum_{s=1}^{\infty} \lambda_s \left(Z_s^2 - 1\right), \tag{A.2}$$

where $\{Z_s : s \geqslant 1\}$ *are i.i.d.* $\mathcal{N}(0,1)$ *and* $\{\lambda_s\}_{s\geqslant 1}$ *are the eigenvalues (with repetitions) of the Hilbert-Schimdt operator* $\mathcal{H}_{\mathsf{H}^\circ}$ *defined as:*

$$\mathcal{H}_{\mathsf{H}^\circ}[f(x)] = \int_{\mathcal{X}} \mathsf{H}^\circ(x,y) f(y) \mathrm{d}P(y), \tag{A.3}$$

with $\mathsf{H}^\circ(x,y) := \mathsf{H}(x,y) - \mathbb{E}_{X\sim P}\mathsf{H}(X,y) - \mathbb{E}_{X'\sim P}\mathsf{H}(x,X') + \mathbb{E}_{X,X'\sim P}\mathsf{H}(X,X')$. *Moreover, the characteristic function of* $Z(\mathsf{H})$ *at* $t \in \mathbb{R}$ *is given by:*

$$\Phi_{Z(\mathsf{H})}(t) := \mathbb{E}\left[e^{\iota t Z(\mathsf{H})}\right] = \prod_{s=1}^{\infty} \frac{e^{-\frac{\iota\lambda_s t}{\rho(1-\rho)}}}{\sqrt{1 - \frac{2\iota\lambda_s t}{\rho(1-\rho)}}}. \tag{A.4}$$

Remark A.1. The convergence in (A.2) is a consequence of [23, Theorem 5], while the expression of the characteristic function in (A.4) follows from [31, Proposition 6.1]. (Note that Proposition A.1 also follows from the proof of Proposition C.1 in Section C by setting $h = 0$).

We now present the proof of Theorem 3.1 in Appendix A.1. The proof of Corollary 3.1 is given in Appendix A.2.

A.1. **Proof of Theorem 3.1.** First recall the definition of $\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$ from (2.10). Note that for $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_r)^\top \in \mathbb{R}^r$,

$$\boldsymbol{\eta}^\top \mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] = \sum_{a=1}^{r} \eta_a \mathrm{MMD}^2[\mathsf{K}_a, \mathscr{X}_m, \mathscr{Y}_n] = \mathrm{MMD}^2[\mathsf{H}_{\boldsymbol{\eta}}, \mathscr{X}_m, \mathscr{Y}_n], \tag{A.5}$$

where $\mathsf{H}_{\boldsymbol{\eta}} := \sum_{a=1}^r \eta_a \mathsf{K}_a$. Clearly, $\mathsf{H}_{\boldsymbol{\eta}}$ is a measurable and symmetric function and $\mathsf{H}_{\boldsymbol{\eta}} \in L^2\left(\mathcal{X}, P^2\right)$ (by Assumption 2.1). Then by Proposition A.1,

$$Z_{m,n}(\mathsf{H}_{\boldsymbol{\eta}}) := (m+n)\mathrm{MMD}^2\left[\mathsf{H}_{\boldsymbol{\eta}}, \mathscr{X}_m, \mathscr{Y}_n\right] \xrightarrow{D} Z(\mathsf{H}_{\boldsymbol{\eta}}) = \frac{1}{\rho(1-\rho)} \sum_{\lambda\in\Lambda(\boldsymbol{\eta})} \lambda\left(Z_\lambda^2 - 1\right). \tag{A.6}$$

where $\{Z_\lambda\}_{\lambda\in\Lambda(\boldsymbol{\eta})}$ are i.i.d. $\mathcal{N}(0,1)$ and $\Lambda(\boldsymbol{\eta})$ are the eigenvalues (with repetitions) of the Hilbert-Schimdt operator:

$$\mathcal{H}_{\mathsf{H}^\circ_{\boldsymbol{\eta}}}[f(x)] = \int \mathsf{H}^\circ_{\boldsymbol{\eta}}(x,y) f(y) \mathrm{d}P(y), \tag{A.7}$$

with

$$\begin{aligned}\mathsf{H}^\circ_{\boldsymbol{\eta}}(x,y) &:= \mathsf{H}_{\boldsymbol{\eta}}(x,y) - \mathbb{E}_{X\sim P}\mathsf{H}_{\boldsymbol{\eta}}(X,y) - \mathbb{E}_{X'\sim P}\mathsf{H}_{\boldsymbol{\eta}}(x,X') + \mathbb{E}_{X,X'\sim P}\mathsf{H}_{\boldsymbol{\eta}}(X,X') \\ &= \sum_{a=1}^{r} \eta_a \mathsf{K}^\circ_a(x,y),\end{aligned} \tag{A.8}$$

Hence, the operator $\mathcal{H}_{\mathsf{H}^\circ_{\boldsymbol{\eta}}}$ in (A.7) is same as the operator $\mathcal{H}_{\mathcal{K},\boldsymbol{\eta}}$ defined in (3.6). Then Proposition A.1, (A.6), and (A.8) implies,

$$\mathbb{E}\left[e^{\iota Z_{m,n}(\mathsf{H}_{\boldsymbol{\eta}})}\right] \to \mathbb{E}\left[e^{\iota Z(\mathsf{H}_{\boldsymbol{\eta}})}\right] = \prod_{\lambda\in\Lambda(\boldsymbol{\eta})} \frac{e^{-\frac{\iota\lambda}{\rho(1-\rho)}}}{\sqrt{1 - \frac{2\iota\lambda}{\rho(1-\rho)}}} = \Phi(\boldsymbol{\eta}), \tag{A.9}$$

where $\Phi(\boldsymbol{\eta})$ is as defined in (3.5). Since $\boldsymbol{\eta} \in \mathbb{R}^r$ was chosen arbitrarily and $\Phi(\boldsymbol{\eta})$ is continuous at $\boldsymbol{\eta} = \mathbf{0} \in \mathbb{R}^r$ (by Lemma F.1), Levy's continuity theorem [17, Theorem 3.3.17] implies that there exists a random variable $Z_{\mathcal{K}}$ with characteristic function $\Phi(\boldsymbol{\eta})$ such that,

$$\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{D} Z_{\mathcal{K}}. \tag{A.10}$$

We now show that the limit $Z_{\mathcal{K}}$ in (A.10) can be expressed as $G_{\mathcal{K}}$ in (3.4). Towards this, note that by the linearity of multiple stochastic integrals,

$$\mathbb{E}\left[e^{\iota \boldsymbol{\eta}^\top G_{\mathcal{K}}}\right] = \mathbb{E}\left[e^{\frac{\iota}{\rho(1-\rho)}\sum_{a=1}^r \eta_a I_2(\mathsf{K}_{\mathsf{a}}^\circ)}\right] = \mathbb{E}\left[e^{\frac{\iota}{\rho(1-\rho)} I_2(\sum_{a=1}^r \eta_a \mathsf{K}_{\mathsf{a}}^\circ)}\right] = \mathbb{E}\left[e^{\frac{\iota}{\rho(1-\rho)} I_2(\mathsf{H}_{\boldsymbol{\eta}}^\circ)}\right], \tag{A.11}$$

where the last step uses (A.8). Then by [31, Theorem 6.1], $I_2(\mathsf{H}_{\boldsymbol{\eta}}^\circ)$ has characteristic function:

$$\Phi_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}(s) := \mathbb{E}\left[e^{\iota s I_2(\mathsf{H}_{\boldsymbol{\eta}}^\circ)}\right] = \prod_{\lambda \in \Lambda(\boldsymbol{\eta})} \frac{e^{-\iota\lambda s}}{\sqrt{1 - 2\iota\lambda s}} \tag{A.12}$$

where $\Lambda(\boldsymbol{\eta})$ is the set of eigenvalues (with repetition) of the bilinear form $\mathcal{B}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ} : L^2(\mathcal{X}) \times L^2(\mathcal{X}) \to \mathbb{R}$:

$$\mathcal{B}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}(f_P, f_2) := \frac{1}{2}\mathbb{E}\left[I_2(\mathsf{H}_{\boldsymbol{\eta}}^\circ) I_1(f_1) I_1(f_2)\right],$$

for any $f_1, f_2 \in L^2(\mathcal{X}, P)$. Now, using the multiplication formula for stochastic integrals (cf. [31, Theorem 7.33]) gives,

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[I_2(\mathsf{H}_{\boldsymbol{\eta}}^\circ) I_1(f_1) I_1(f_2)\right] &= \frac{1}{2}\int_{\mathcal{X}^2} \left[\mathsf{H}_{\boldsymbol{\eta}}^\circ(x,y) f_1(x) f_2(y) + \mathsf{H}_{\boldsymbol{\eta}}^\circ(x,y) f_1(y) f_2(x)\right] \mathrm{d}P(x)\mathrm{d}P(y) \\ &= \int_{\mathcal{X}^2} \mathsf{H}_{\boldsymbol{\eta}}^\circ(x,y) f_1(x) f_2(y) \mathrm{d}P(x)\mathrm{d}P(y),\end{aligned}$$

where the last equality follows by symmetry of the function $\mathsf{H}_{\boldsymbol{\eta}}^\circ$. This shows that the bilinear form $\mathcal{B}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}$ has the same set of eigenvalues (with repetitions) as that of the operator $\mathcal{H}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}$ defined in (A.8). Hence, combining (A.11) and (A.12) it follows that,

$$\mathbb{E}\left[e^{\iota \boldsymbol{\eta}^\top G_{\mathcal{K}}}\right] = \Phi_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}\left(\frac{1}{\rho(1-\rho)}\right) = \prod_{\lambda \in \Lambda(\boldsymbol{\eta})} \frac{e^{-\frac{\iota\lambda}{\rho(1-\rho)}}}{\sqrt{1 - \frac{2\iota\lambda}{\rho(1-\rho)}}}, \tag{A.13}$$

where $\Lambda(\boldsymbol{\eta})$ is the set of eigenvalues (with repetitions) of the operator $\mathcal{H}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}$, which the same as the operator $\mathcal{H}_{\mathcal{K},\boldsymbol{\eta}}$ (by (3.6)). Note that the RHS of (A.13) equals the function $\Phi(\boldsymbol{\eta})$ defined in (3.5), which implies that $Z_{\mathcal{K}}$ in (A.10) has the same distribution as $G_{\mathcal{K}}$ in (3.4). This completes the proof of Theorem 3.1. □

A.2. **Proof of Corollary 3.1.** Recall the definition of the centered kernel K_a° from (2.13). Then it is easy to see that

$$\mathrm{MMD}^2\left[\mathsf{K}_a, \mathscr{X}_m, \mathscr{Y}_n\right] = \mathrm{MMD}^2\left[\mathsf{K}_a^\circ, \mathscr{X}_m, \mathscr{Y}_n\right].$$

Observe that $\mathbb{E}\left[\mathsf{K}_a^\circ(X_1, X_2) | X_1\right] = 0$ and hence $\mathrm{Cov}\left(\mathsf{K}_a^\circ(X_1, X_2)\mathsf{K}_b^\circ(X_1, X_3)\right) = 0$, for X_1, X_2, X_3 i.i.d. from the distribution P. Then by a direct computation it follows that, for $1 \leqslant a, b \leqslant r$,

$$\sigma_{ab} = \frac{2}{\rho^2(1-\rho)^2}\mathbb{E}_{X,X' \sim P}\left[\mathsf{K}_a^\circ(X, X')\mathsf{K}_b^\circ(X, X')\right]. \tag{A.14}$$

This proves (3.9).

Next, we prove (3.10). For all $1 \leqslant a \leqslant r$, define

$$\mathbf{K}_a^\circ = \left((\mathsf{K}_a^\circ(X_i, X_j)/m)\right)_{1 \leqslant i,j \leqslant m}$$

where K_a° is defined in Theorem 3.1. Also, recall the definition of $\hat{\mathbf{K}}_a^\circ$, for $1 \leqslant a \leqslant r$, from (4.1). Now, observe that for any $1 \leqslant a, b \leqslant r$,

$$\begin{aligned}\left|\mathrm{Tr}\left[\hat{\mathbf{K}}_a^\circ \hat{\mathbf{K}}_b^\circ\right] - \mathrm{Tr}\left[\mathbf{K}_a^\circ \mathbf{K}_b^\circ\right]\right| &\leqslant \left|\mathrm{Tr}\left[\hat{\mathbf{K}}_a^\circ \hat{\mathbf{K}}_b^\circ\right] - \mathrm{Tr}\left[\mathbf{K}_a^\circ \hat{\mathbf{K}}_b^\circ\right]\right| + \left|\mathrm{Tr}\left[\mathbf{K}_a^\circ \hat{\mathbf{K}}_b^\circ\right] - \mathrm{Tr}\left[\mathbf{K}_a^\circ \mathbf{K}_b^\circ\right]\right| \\ &\leqslant \|\hat{\mathbf{K}}_a^\circ\|\|\hat{\mathbf{K}}_a^\circ - \mathbf{K}_a^\circ\| + \|\mathbf{K}_a^\circ\|\|\hat{\mathbf{K}}_b^\circ - \mathbf{K}_b^\circ\| \qquad \text{(A.15)}\end{aligned}$$

Since $\mathsf{K}_a^\circ \in L^2(\mathcal{X}^2, P^2)$, by the strong law of large number for U-statistics (see [57, Theorem 5.4.A])

$$\|\mathbf{K}_a^\circ\|^2 = \frac{1}{m^2}\sum_{1\leqslant i,j\leqslant m} \mathsf{K}_a^\circ(X_i, X_j)^2 \overset{a.s.}{\to} \mathbb{E}_{X,X'\sim P}[\mathsf{K}_a^\circ(X, X')].$$

Also, following the proof of Lemma B.4 we have $\|\hat{\mathbf{K}}_a^\circ - \mathbf{K}_a^\circ\| \overset{a.s.}{\to} 0$. This implies, $\|\hat{\mathbf{K}}_a^\circ\|^2 \overset{a.s.}{\to} \mathbb{E}_{X,X'\sim P}[\mathsf{K}_a^\circ(X, X')]$. Thus, combining the above conclusions with (A.15) gives,

$$\left|\mathrm{Tr}\left[\hat{\mathbf{K}}_a^\circ \hat{\mathbf{K}}_b^\circ\right] - \mathrm{Tr}\left[\mathbf{K}_a^\circ \mathbf{K}_b^\circ\right]\right| \overset{a.s.}{\to} 0 \qquad \text{(A.16)}$$

Since, for all $1 \leqslant a \leqslant r$, $\mathsf{K}_a^\circ \in L^2(\mathcal{X}^2, P^2)$, then by [57, Theorem 5.4.A],

$$\begin{aligned}\mathrm{Tr}\left[\mathbf{K}_a^\circ \mathbf{K}_b^\circ\right] = \frac{1}{m^2}\sum_{1\leqslant i,j\leqslant m} \mathsf{K}_a^\circ(X_i, X_j)\mathsf{K}_b^\circ(X_i, X_j) &\overset{a.s.}{\to} \mathbb{E}_{X\sim P}\left[\mathsf{K}_a^\circ(X_1, X_2)\mathsf{K}_b^\circ(X_1, X_2)\right] \\ &= \frac{\rho^2(1-\rho)^2}{2}\sigma_{ab}, \qquad \text{(A.17)}\end{aligned}$$

where the last equality follows from (A.14). Now, recalling (2.14) and applying (A.16) and (A.17) note that

$$\begin{aligned}\hat{\sigma}_{ab} = \frac{2}{\hat{\rho}^2(1-\hat{\rho})^2}\cdot\frac{1}{m^2}\sum_{1\leqslant i,j\leqslant m} \hat{\mathsf{K}}_a^\circ(X_i, X_j)\hat{\mathsf{K}}_b^\circ(X_i, X_j) &= \frac{2}{\hat{\rho}^2(1-\hat{\rho})^2}\,\mathrm{Tr}\left[\hat{\mathbf{K}}_a^\circ \hat{\mathbf{K}}_b^\circ\right] \\ &\overset{a.s.}{\to} \sigma_{ab}.\end{aligned}$$

This completes the proof of (3.10). The result in (3.11) follows from (3.10) combined with Theorem 3.1, Slutsky's theorem, and the continuous mapping theorem.

Appendix B. Proof of Theorem 4.1

Suppose $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ be a collection of r characteristic kernels satisfying the conditions of Theorem 4.1. We will prove Theorem 4.1 by showing that every linear combination of $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)$ converges to the corresponding linear combination of $G_{\mathcal{K}}$. To this end, suppose $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_r)^\top \in \mathbb{R}^r$ and define $\mathsf{H}_{\boldsymbol{\eta}} = \sum_{a=1}^r \eta_a \mathsf{K}_a$. Let $\mathsf{H}_{\boldsymbol{\eta}}^\circ = \sum_{a=1}^r \eta_a \mathsf{K}_a^\circ$ be as defined in (A.8) and $\{\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)\}_{s\geqslant 1}$ be the eigenvalues of the operator $\mathcal{H}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}$ as in (A.7). Also, define

$$\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ = \boldsymbol{C}\hat{\mathbf{H}}_{\boldsymbol{\eta}}\boldsymbol{C}/m,$$

where $\hat{\mathbf{H}}_{\boldsymbol{\eta}} = (\mathsf{H}_{\boldsymbol{\eta}}(X_i, X_j))_{1\leqslant i,j\leqslant m}$ and $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ is the centering matrix as defined in (4.1). Note that

$$\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ = \left(\left(\frac{\hat{\mathsf{H}}_{\boldsymbol{\eta}}^\circ(X_i, X_j)}{m}\right)\right)_{1\leqslant i,j\leqslant m},$$

where, similar to (2.15),

$$\hat{\mathsf{H}}_{\boldsymbol{\eta}}^\circ(x, y) = \mathsf{H}_{\boldsymbol{\eta}}(x, y) - \frac{1}{m}\sum_{u=1}^m \mathsf{H}_{\boldsymbol{\eta}}(X_u, y) - \frac{1}{m}\sum_{v=1}^m \mathsf{H}_{\boldsymbol{\eta}}(x, X_v) + \frac{1}{m^2}\sum_{1\leqslant u,v\leqslant m} \mathsf{H}_{\boldsymbol{\eta}}(X_u, X_v). \qquad \text{(B.1)}$$

Recalling the definition of $\hat{\mathbf{K}}_a^\circ$ from (4.1), observe that

$$\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ = \sum_{a=1}^{r} \eta_a \hat{\mathbf{K}}_a^\circ. \tag{B.2}$$

Let $\{\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)\}_{1\leqslant s\leqslant m}$ be the eigenvalues of the matrix $\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ$. Recall that $\gamma = \frac{1}{\rho(1-\rho)}$. Then we have the following proposition:

Proposition B.1. *Suppose $\{\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)\}_{s\geqslant 1}$ and $\{\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)\}_{1\leqslant s\leqslant m}$ be as defined above. Then there exists a set $\mathcal{Q}_0 \in \mathscr{B}(\mathcal{X})$ (not depending on $\boldsymbol{\eta}$) with $\mathbb{P}(\mathcal{Q}_0) = 1$ such that on the set $\mathcal{Q}_0$,*

$$\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)\left(W_s^2 - \gamma\right) | \mathscr{X}_m \xrightarrow{D} \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)\left(Z_s^2 - \gamma\right), \tag{B.3}$$

as $m \to \infty$, where $\{W_s, Z_s : s \geqslant 1\}$ are i.i.d. from $\mathcal{N}(0,\gamma)$ independent of $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$.

The proof of Proposition B.1 is given in Appendix B.1. We first show how this can be used to complete the proof of Theorem 4.1. To this end, note that by definition, for all $s \geqslant 1$, $W_s = \sqrt{\gamma} W_s^\circ$, where $\{W_s^\circ : s \geqslant 1\}$ are i.i.d. from $\mathcal{N}(0,1)$. Define, for $s \geqslant 1$,

$$\hat{W}_s = \sqrt{\hat{\gamma}} W_s^\circ,$$

where $\hat{\gamma} := \frac{1}{\hat{\rho}(1-\hat{\rho})} = \frac{mn}{(m+n)^2}$. Now observe that by Lemma B.1 on $\mathcal{Q} := \mathcal{Q}_1 \bigcap \mathcal{Q}_0$,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)(W_s^2 - \gamma) - \sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)(\hat{W}_s^2 - \hat{\gamma})^2\right)^2 \middle| \mathscr{X}_m\right] &= \sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)^2(\gamma - \hat{\gamma})^2 \\ &= (\gamma - \hat{\gamma})^2 \|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ\|^2 \to 0,\end{aligned}$$

and hence on the set $\mathcal{Q}$,

$$\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)\left(\hat{W}_s^2 - \hat{\gamma}\right) | \mathscr{X}_m \xrightarrow{D} \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)\left(Z_s^2 - \gamma\right). \tag{B.4}$$

By the spectral decomposition,

$$\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ = \boldsymbol{Q}_m \boldsymbol{\Lambda}_m \boldsymbol{Q}_m^\top,$$

where $\boldsymbol{\Lambda}_m = \mathrm{diag}(\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ))_{1\leqslant s\leqslant m}$ and $\boldsymbol{Q}_m^\top \boldsymbol{Q}_m = \boldsymbol{Q}_m \boldsymbol{Q}_m^\top = \boldsymbol{I}$ is an orthogonal matrix. Observe that for $\boldsymbol{W}_m = (\hat{W}_1, \ldots, \hat{W}_m)^\top$,

$$\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)\left(\hat{W}_s^2 - \hat{\gamma}\right) = \boldsymbol{W}_m^\top \boldsymbol{\Lambda}_m \boldsymbol{W}_m - \hat{\gamma}\,\mathrm{Tr}[\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ] = \boldsymbol{Z}_m^\top \hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ \boldsymbol{Z}_m - \hat{\gamma}\,\mathrm{Tr}[\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ], \tag{B.5}$$

where $\boldsymbol{Z}_m = \boldsymbol{Q}_m \boldsymbol{W}_m \sim \mathcal{N}_m(\boldsymbol{0}, \hat{\gamma}\boldsymbol{I})$ is independent of $\mathscr{X}_m$. This is because $\boldsymbol{Q}_m$ is orthogonal and, hence, $\boldsymbol{Z}_m | \mathscr{X}_m \sim \mathcal{N}_m(\boldsymbol{0}, \hat{\gamma}\boldsymbol{I})$, which implies $\boldsymbol{Z}_m \sim \mathcal{N}_m(\boldsymbol{0}, \hat{\gamma}\boldsymbol{I})$. By (B.2), (B.5), and recalling the definition of $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)$ from (4.2) it follows that

$$\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^\circ)\left(\hat{W}_s^2 - \hat{\gamma}\right) = \boldsymbol{\eta}^\top \mathcal{E}(\mathcal{K}, \mathscr{X}_m). \tag{B.6}$$

Hence, by (B.4) on a set $\mathcal{Q}$ with $\mathbb{P}(\mathcal{Q}) = 1$, as $m \to \infty$,

$$\boldsymbol{\eta}^\top \mathcal{E}(\mathcal{K}, \mathscr{X}_m) | \mathscr{X}_m \xrightarrow{D} \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)\left(Z_s^2 - \gamma\right) \overset{D}{=} Z(\mathsf{H}_{\boldsymbol{\eta}}), \tag{B.7}$$

where the last step uses (A.6). From (A.5) and (A.6) we know that $Z(\mathsf{H}_{\boldsymbol{\eta}})$ is the limiting distribution of $\boldsymbol{\eta}^\top \mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$. Hence, by Theorem 3.1, $\sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^\circ_{\boldsymbol{\eta}}) \left(Z_s^2 - \gamma\right) \stackrel{D}{=} \boldsymbol{\eta}^\top G_{\mathcal{K}}$. This implies, by (B.6) and (B.7), on the set $\mathcal{Q}$,

$$\boldsymbol{\eta}^\top \mathcal{E}(\mathcal{K}, \mathscr{X}_m)|\mathscr{X}_m \xrightarrow{D} \boldsymbol{\eta}^\top G_{\mathcal{K}},$$

as $m \to \infty$. Hence, by the Cramer-Wold device, on the set $\mathcal{Q}$, $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)|\mathscr{X}_m \xrightarrow{D} G_{\mathcal{K}}$, as $m \to \infty$. This completes the proof of Theorem 4.1. □

B.1. **Proof of Proposition B.1.** The proof of Proposition B.1 is organized as follows. First we show that the L_2 norm of the matrix $\hat{\mathbf{H}}^\circ_{\boldsymbol{\eta}}$ converges to the L_2 norm of the operator $\mathcal{H}_{\mathsf{H}^\circ_{\boldsymbol{\eta}}}$ almost surely (proof is given in Appendix B.1.1).

Lemma B.1. *There exists a set $\mathcal{Q}_1 \in \mathscr{B}(\mathcal{X})$ (not depending on $\boldsymbol{\eta}$) with $\mathbb{P}(\mathcal{Q}_1) = 1$ such that on $\mathcal{Q}_1$,*

$$\|\hat{\mathbf{H}}^\circ_{\boldsymbol{\eta}}\| \to \|\mathsf{H}^\circ_{\boldsymbol{\eta}}\|, \tag{B.8}$$

as $m \to \infty$.

Next, we show that the ℓ-th moment (power sum) of the eigenvalues of $\hat{\mathbf{H}}^\circ_{\boldsymbol{\eta}}$ converges to the ℓ-th moment (power sum) of the eigenvalues of $\mathcal{H}_{\mathsf{H}^\circ_{\boldsymbol{\eta}}}$, for $\ell \geqslant 3$, almost surely (proof is given in Appendix B.1.2).

Lemma B.2. *There exists a set $\mathcal{Q}_2 \in \mathscr{B}(\mathcal{X})$ (not depending on $\boldsymbol{\eta}$) with $\mathbb{P}(\mathcal{Q}_2) = 1$ such that on $\mathcal{Q}_2$,*

$$\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}^\circ_{\boldsymbol{\eta}})^\ell \to \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^\circ_{\boldsymbol{\eta}})^\ell, \tag{B.9}$$

for all $\ell \geqslant 3$, as $m \to \infty$.

Finally, we show that if (B.8) and (B.9) holds, then the convergence in (B.3) holds (proof is given in Appendix B.1.3).

Lemma B.3. *Suppose (B.8) and (B.9) holds. Then on the set $\mathcal{Q}_0 = \mathcal{Q}_1 \bigcap \mathcal{Q}_2$, the convergence in (B.3) holds.*

Since $\mathbb{P}(\mathcal{Q}_0) = 1$ and $\mathcal{Q}_0$ does not depend on $\boldsymbol{\eta}$, the above 3 lemmas combined completes the proof of Proposition B.1.

B.1.1. *Proof of Lemma B.1.* Define

$$\mathbf{H}^\circ_{\boldsymbol{\eta}} = \left(\left(\frac{\mathsf{H}^\circ_{\boldsymbol{\eta}}(X_i, X_j)}{m} \right) \right)_{1 \leqslant i,j \leqslant m},$$

where

$$\mathsf{H}^\circ_{\boldsymbol{\eta}}(X_i, X_j) := \mathsf{H}_{\boldsymbol{\eta}}(X_i, X_j) - \mathbb{E}_{X \sim P} \mathsf{H}_{\boldsymbol{\eta}}(X, X_j) - \mathbb{E}_{X' \sim P} \mathsf{H}_{\boldsymbol{\eta}}(X_i, X') + \mathbb{E}_{X, X' \sim P} \mathsf{H}_{\boldsymbol{\eta}}(X, X'), \tag{B.10}$$

First we will show that $\mathbf{H}^\circ_{\boldsymbol{\eta}}$ and $\hat{\mathbf{H}}^\circ_{\boldsymbol{\eta}}$ are asymptotically close.

Lemma B.4. *There exists a set $\mathcal{R}_1 \in \mathscr{B}(\mathcal{X})$ (not depending on $\boldsymbol{\eta}$) with $\mathbb{P}(\mathcal{R}_1) = 1$ such that*

$$\lim_{m \to \infty} \|\hat{\mathbf{H}}^\circ_{\boldsymbol{\eta}} - \mathbf{H}^\circ_{\boldsymbol{\eta}}\|^2 = 0. \tag{B.11}$$

Proof. Note that

$$\|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ} - \mathbf{H}_{\boldsymbol{\eta}}^{\circ}\|^2 = \frac{1}{m^2} \sum_{1 \leqslant i,j \leqslant m} \left(\hat{\mathsf{H}}_{\boldsymbol{\eta}}^{\circ}(X_i, X_j) - \mathsf{H}_{\boldsymbol{\eta}}^{\circ}(X_i, X_j)\right)^2.$$

Hence, by (B.1) and (B.10),

$$\|\mathbf{H}_{\boldsymbol{\eta}}^{\circ} - \hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}\|^2 \leqslant 3(T_1 + T_2 + T_3), \tag{B.12}$$

where

$$T_1 := \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{m} \sum_{v=1}^{m} \mathsf{H}_{\boldsymbol{\eta}}(X_i, X_v) - \mathbb{E}_{X' \sim P}[\mathsf{H}_{\boldsymbol{\eta}}(X_i, X')] \right)^2,$$

$$T_2 := \frac{1}{m} \sum_{j=1}^{m} \left(\frac{1}{m} \sum_{u=1}^{m} \mathsf{H}_{\boldsymbol{\eta}}(X_u, X_j) - \mathbb{E}_{X' \sim P}[\mathsf{H}_{\boldsymbol{\eta}}(X', X_j)] \right)^2,$$

$$T_3 := \left(\frac{1}{m^2} \sum_{1 \leqslant u,v \leqslant m} \mathsf{H}_{\boldsymbol{\eta}}(X_u, X_v) - \mathbb{E}_{X,X' \sim P}[\mathsf{H}_{\boldsymbol{\eta}}(X, X')] \right)^2.$$

Since $\mathsf{H}_{\boldsymbol{\eta}} = \sum_{a=1}^{r} \eta_a \mathsf{K}_a$,

$$T_1 \leqslant C_1^{(r)} \sum_{a=1}^{r} \eta_a^2 \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{m} \sum_{v=1}^{m} \mathsf{K}_a(X_i, X_v) - \mathbb{E}_{X' \sim P}[\mathsf{K}_a(X_i, X')] \right)^2 \right\}. \tag{B.13}$$

where $C_1^{(r)} > 0$ is a constant depending on r only. By Lemma F.3, for every $1 \leqslant a \leqslant r$ there exists a set $\mathcal{B}_{\mathsf{K}_a} \in \mathscr{B}(\mathcal{X})$ with $\mathbb{P}(\mathcal{B}_{\mathsf{K}_a}) = 1$ such that

$$\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{m} \sum_{v=1}^{m} \mathsf{K}_a(X_i, X_v) - \mathbb{E}_{X' \sim P}[\mathsf{K}_a(X_i, X')] \right)^2 = 0.$$

Define $\mathcal{B}_{\mathcal{K}} = \bigcap_{s=1}^{r} \mathcal{B}_{\mathsf{K}_a}$. Clearly, $\mathbb{P}(\mathcal{B}_{\mathcal{K}}) = 1$. By Lemma F.3 and (B.13), on the set $\mathcal{B}_{\mathcal{K}}$, $\lim_{m \to \infty} T_1 = 0$. Similarly, on the set $\mathcal{B}_{\mathcal{K}}$, $\lim_{m \to \infty} T_2 = 0$.

Also, by [57, Theorem 5.4.A] there is a set $\mathcal{E}_{\mathcal{K}}$ (depending $\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r$, but not on $\boldsymbol{\eta}$) with $\mathbb{P}(\mathcal{E}_{\mathcal{K}}) = 1$ such that on the set $\mathcal{E}_{\mathcal{K}}$, for all $1 \leqslant a \leqslant r$,

$$\lim_{m \to \infty} \frac{1}{m^2} \sum_{1 \leqslant u,v \leqslant m} \mathsf{K}_a(X_u, X_v) - \mathbb{E}_{X,X' \sim P}[\mathsf{K}_a(X, X')] = 0.$$

Then on the set $\mathcal{E}_{\mathcal{K}}$,

$$T_3 \leqslant C_2^{(r)} \sum_{a=1}^{r} \eta_a^2 \left(\left(\frac{1}{m^2} \sum_{1 \leqslant u,v \leqslant m} \mathsf{K}_a(X_u, X_v) - \mathbb{E}_{X,X' \sim P}[\mathsf{K}_a(X, X')] \right)^2 \right) \to 0, \tag{B.14}$$

where $C_2^{(r)} > 0$ is a constant depending on r only. Combining (B.13) and (B.14) with (B.12) shows that (B.11) holds on $\mathcal{R}_1 = \mathcal{B}_{\mathcal{K}} \bigcap \mathcal{E}_{\mathcal{K}}$. Since $\mathbb{P}(\mathcal{R}_1) = 1$, this completes the proof of Lemma B.4. □

Now, we compute $\|\mathbf{H}_{\boldsymbol{\eta}}^{\circ}\|$. By [57, Theorem 5.4.A], there exists a set $\mathcal{R}_2 \in \mathscr{B}(\mathcal{X})$ with $\mathbb{P}(\mathcal{R}_2) = 1$ such that on the set $\mathcal{R}_2$,

$$\|\mathbf{H}_{\boldsymbol{\eta}}^{\circ}\|^2 = \frac{1}{m^2} \sum_{1 \leqslant i,j \leqslant n} \mathsf{H}_{\boldsymbol{\eta}}^{\circ}(X_i, X_j)^2$$

$$= \sum_{a=1}^{r} \eta_a^2 \left(\frac{1}{m^2} \sum_{1 \leqslant i,j \leqslant n} \mathsf{K}_a^\circ(X_i, X_j)^2 \right) + \sum_{1 \leqslant s \neq t \leqslant r} \eta_a \eta_b \left(\frac{1}{m^2} \sum_{1 \leqslant i,j \leqslant n} \mathsf{K}_a^\circ(X_i, X_j) \mathsf{K}_b^\circ(X_i, X_j) \right)$$

$$\to \sum_{a=1}^{r} \eta_a^2 \mathbb{E}_{X,X' \sim P}[\mathsf{K}_a^\circ(X, X')^2] + \sum_{1 \leqslant a \neq b \leqslant r} \eta_a \eta_b \mathbb{E}_{X,X' \sim P}[\mathsf{K}_a^\circ(X, X') \mathsf{K}_b^\circ(X, X')]$$

$$= \mathbb{E}\left[\left(\sum_{a=1}^{r} \eta_a \mathsf{K}_a^\circ(X, X') \right)^2 \right]$$

$$= \mathbb{E}_{X,X' \sim P}\left[\mathsf{H}_{\boldsymbol{\eta}}^\circ(X, X')^2 \right] = \left\| \mathsf{H}_{\boldsymbol{\eta}}^\circ \right\|^2. \tag{B.15}$$

Combining the above together with Lemma B.4 and choosing $\mathcal{Q}_1 = \mathcal{R}_1 \bigcap \mathcal{R}_2$, the result in Lemma B.1 follows.

B.1.2. *Proof of Lemma B.2.* Define

$$\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-} = \left(\left((1 - \delta_{ij}) \frac{\mathsf{H}_{\boldsymbol{\eta}}^\circ(X_i, X_j)}{m} \right) \right)_{1 \leqslant i,j \leqslant m}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. First, we will show that the ℓ-th moment (power sums) of the eigenvalues of $\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-}$ converges to the ℓ-th moment of the eigenvalues of $\mathcal{H}_{\mathsf{H}_{\boldsymbol{\eta}}^\circ}$, for $\ell \geqslant 3$, almost surely.

Lemma B.5. *Suppose $\{\lambda_s(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-}) : 1 \leqslant s \leqslant m\}$ are the eigenvalues of $\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-}$. Then there exists a set $\mathcal{E}_1 \in \mathscr{B}(\mathcal{X})$ with $\mathbb{P}(\mathcal{E}_1) = 1$ such that on $\mathcal{E}$*

$$\lim_{m \to \infty} \sum_{s=1}^{m} \lambda_s(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-})^\ell = \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)^\ell,$$

for all $\ell \geqslant 3$.

Proof. For fixed $\ell \geqslant 3$, observe that,

$$\sum_{s=1}^{m} \lambda_s(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-})^\ell = \operatorname{tr}\left[\left(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-} \right)^\ell \right]$$

$$= \frac{1}{m^\ell} \sum_{1 \leqslant i_1 \neq i_2 \neq \cdots \neq i_\ell \leqslant m} \mathsf{H}_{\boldsymbol{\eta}}^\circ(X_{i_1}, X_{i_2}) \mathsf{H}_{\boldsymbol{\eta}}^\circ(X_{i_2}, X_{i_3}) \cdots \mathsf{H}_{\boldsymbol{\eta}}^\circ(X_{i_\ell}, X_{i_1}). \tag{B.16}$$

Recalling $\mathsf{H}_{\boldsymbol{\eta}}^\circ = \sum_{a=1}^{r} \eta_a \mathsf{K}_a^\circ$, (B.16) can be written as:

$$\sum_{s=1}^{m} \lambda_s(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-})^\ell = \sum_{j_1, \cdots, j_\ell \in \{1,2,\ldots,r\}} \prod_{t=1}^{\ell} \eta_{j_t} \left(\frac{1}{m^\ell} \sum_{1 \leqslant i_1 \neq \cdots \neq i_\ell \leqslant m} \prod_{t=1}^{\ell} \mathsf{K}_{j_t}^\circ(X_{i_t}, X_{i_{t+1}}) \right). \tag{B.17}$$

Now, since $\mathsf{H}_{\boldsymbol{\eta}}^\circ \in L^2(\mathcal{X}^2, P^2)$, by the spectral theorem

$$\mathsf{H}_{\boldsymbol{\eta}}^\circ(x, y) = \sum_{s=1}^{\infty} \lambda_s \phi_s(x) \phi_s(y), \tag{B.18}$$

where $\{\lambda_1, \lambda_2, \ldots\}$ are the eigenvalues of $\mathsf{H}_{\boldsymbol{\eta}}^\circ$ and $\{\phi_1, \phi_2, \ldots\}$ are the corresponding eigenvectors which form an orthonormal basis of $L^2(\mathcal{X}, P)$. Using the spectral representation in (B.18) and the orthonormality of the eigenvectors it follows that

$$\sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^\circ)^\ell = \int \mathsf{H}_{\boldsymbol{\eta}}^\circ(x_1, x_2) \mathsf{H}_{\boldsymbol{\eta}}^\circ(x_2, x_3) \cdots \mathsf{H}_{\boldsymbol{\eta}}^\circ(x_\ell, x_1) \mathrm{d}P(x_1) \cdots \mathrm{d}P(x_\ell) \tag{B.19}$$

$$= \sum_{j_1,\cdots,j_\ell \in \{1,2,\ldots,r\}} \prod_{t=1}^{\ell} \eta_{j_t} \mathbb{E}\left[\mathsf{K}^{\circ}_{j_1}(X_1, X_2) \cdots \mathsf{K}^{\circ}_{j_\ell}(X_\ell, X_1)\right], \tag{B.20}$$

(Note that the R.H.S. of (B.19) is finite, since $\sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^{\circ}_{\boldsymbol{\eta}})^\ell \leqslant \|\mathsf{H}^{\circ}_{\boldsymbol{\eta}}\|^\ell$, by Finner's inequality [18] (see also [45, Theorem 3.1])) since $\mathsf{H}^{\circ}_{\boldsymbol{\eta}} = \sum_{a=1}^{r} \eta_a \mathsf{K}^{\circ}_a$. Hence, using [57, Theorem 5.4A], (B.17), and (B.20), we can find a set $\mathcal{E}_1 \in \mathscr{B}(\mathcal{X})$ with $\mathbb{P}(\mathcal{E}_1) = 1$ such that on $\mathcal{E}_1$, as $m \to \infty$,

$$\sum_{s=1}^{m} \lambda_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})^\ell \to \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^{\circ}_{\boldsymbol{\eta}})^\ell,$$

for all $\ell \geqslant 3$. This completes the proof of Lemma B.5. □

Now, we will show the (B.9), that is,

$$\lim_{m\to\infty} \sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}})^\ell = \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^{\circ}_{\boldsymbol{\eta}})^\ell,$$

for all $\ell \geqslant 3$, on set $\mathcal{Q}_2$ with $\mathbb{P}(\mathcal{Q}_2) = 1$. To this end, note that for all $\ell \geqslant 3$,

$$\begin{aligned} &\left|\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}})^\ell - \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^{\circ}_{\boldsymbol{\eta}})^\ell\right| \\ &\leqslant \left|\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}})^\ell - \sum_{s=1}^{m} \lambda_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})^\ell\right| + \left|\sum_{s=1}^{m} \lambda_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})^\ell - \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}^{\circ}_{\boldsymbol{\eta}})^\ell\right|. \end{aligned} \tag{B.21}$$

On the set $\mathcal{E}_1$ as in Lemma B.5, the second term above converges to zero as $m \to \infty$. To bound the first term, observe that

$$\left|\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}})^\ell - \sum_{s=1}^{m} \lambda_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})^\ell\right| \leqslant \sum_{s=1}^{\infty} \left|\lambda^+_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}})^\ell - \lambda^+_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})^\ell\right| + \sum_{s=1}^{\infty} \left|\lambda^-_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}})^\ell - \lambda^-_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})^\ell\right|,$$

where

- $\lambda^+_1(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}) \geqslant \lambda^+_2(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}) \geqslant \ldots \geqslant 0$ are the non-negative eigenvalues of $\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}$ (and similarly for $\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}}$) arranged in non-increasing order.
- $\lambda^-_1(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}) \leqslant \lambda^+_2(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}) \leqslant \ldots \leqslant 0$ are the non-positive eigenvalues of $\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}$ (and similarly for $\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}}$) arranged in non-decreasing order.

(Note that we have set $\lambda^{\pm}_s$ to zero whenever appropiate to extend the sequence to infinity.) Now, recalling the definitions of $\mathbf{H}^{\circ}_{\boldsymbol{\eta}}$ and $\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}}$ gives,

$$\|\mathbf{H}^{\circ}_{\boldsymbol{\eta}} - \mathbf{H}^{\circ,-}_{\boldsymbol{\eta}}\| = \frac{1}{m^2} \sum_{i=1}^{m} \mathsf{H}^{\circ}_{\boldsymbol{\eta}}(X_i, X_i)^2 \to 0, \tag{B.22}$$

on a set $\bar{\mathcal{E}}_2 \in \mathcal{B}(\mathcal{X})$. Therefore, recalling Lemma B.4, by (B.22) and the Hoffman-Wielandt inequality [29] (see also [36, Theorem 2.2]),

$$\lim_{m\to\infty} \left(\sum_{s=1}^{\infty} \left(\lambda^+_s(\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}) - \lambda^+_s(\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}})\right)^2\right)^{\frac{1}{2}} \leqslant \lim_{m\to\infty} \left(\|\mathbf{H}^{\circ,-}_{\boldsymbol{\eta}} - \mathbf{H}^{\circ}_{\boldsymbol{\eta}}\| + \|\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}} - \mathbf{H}^{\circ}_{\boldsymbol{\eta}}\|\right) = 0. \tag{B.23}$$

on the set $\mathcal{E}_2 := \bar{\mathcal{E}}_2 \bigcap \mathcal{R}_1$. Moreover, by Lemma B.1, (B.15) and (B.22), on the set $\mathcal{Q}_1$, $\lim_{m\to\infty} \|\hat{\mathbf{H}}^{\circ}_{\boldsymbol{\eta}}\| = \|\mathsf{H}^{\circ}_{\boldsymbol{\eta}}\|$. Hence, by Lemma B.4 on the set $\mathcal{Q}_1 \bigcap \mathcal{E}_2$ there is a constant $C > 0$

such that $\max\{\|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}\|, \|\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-}\|, \|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|\} \leqslant C$. This implies, on the set $\mathcal{Q}_1 \bigcap \mathcal{E}_2$,

$$\max\{|\lambda_s^+(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})|, |\lambda_s^+(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-})|\} \leqslant \frac{C}{\sqrt{s}}, \tag{B.24}$$

for all $s \geqslant 1$. Then using (B.23), (B.24), and the dominated convergence theorem,

$$\lim_{m\to\infty} \sum_{s=1}^{\infty} \left|\lambda_s^+(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^{\ell} - \lambda_s^+(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-})^{\ell}\right| = 0,$$

for all $\ell \geqslant 3$, on the set $\mathcal{Q}_1 \bigcap \mathcal{E}_2$. Similarly, $\lim_{m\to\infty} \sum_{s=1}^{\infty} |\lambda_s^-(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^{\ell} - \lambda_s^-(\mathbf{H}_{\boldsymbol{\eta}}^{\circ,-})^{\ell}| = 0$, for all $\ell \geqslant 3$, on the set $\mathcal{Q}_1 \bigcap \mathcal{E}_2$. Therefore, on the set $\mathcal{Q}_2 := \mathcal{Q}_1 \bigcap \mathcal{E}_1 \bigcap \mathcal{E}_2$, from (B.21),

$$\lim_{m\to\infty} \sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^{\ell} = \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})^{\ell},$$

for all $\ell \geqslant 3$. Since $\mathbb{P}(\mathcal{Q}_2) = 1$, this completes the proof of Lemma B.2.

B.1.3. *Proof of Lemma B.3.* Recall that $\{\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})\}_{1\leqslant s\leqslant m}$ are the eigenvalues of $\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}$. For $s > m$ define $\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}) = 0$. Consider, for all $m \geqslant 1$,

$$Y_m := \sum_{s=1}^{\infty} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})(W_s^2 - \gamma). \tag{B.25}$$

Then by [8, Proposition 7.1] observe that

$$M_{Y_m|\mathscr{X}_m}(t) := \mathbb{E}\left[e^{tY_m}\middle|\mathscr{X}_m\right] = \prod_{s=1}^{m} \frac{\exp\left(-\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t\right)}{\sqrt{1-2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t}}, \text{ for all } |t| < \frac{1}{8\gamma}\left(\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^2\right)^{-\frac{1}{2}}.$$

By definition, $\sum_{s=1}^{m} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^2 = \sum_{s=1}^{\infty} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^2 = \|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}\|^2$. Taking logarithm and expanding gives,

$$\log M_{Y_m|\mathscr{X}_m}(t) = \gamma^2 t^2 \|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}\|^2 + \frac{1}{2}\sum_{k=3}^{\infty}\sum_{s=1}^{m} \frac{(2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t)^k}{k}. \tag{B.26}$$

Also, Denote $Z(\mathsf{H}_{\boldsymbol{\eta}}^{\circ}) = \sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})(Z_s^2 - \gamma)$. Then by Lemma F.2,

$$\begin{aligned}\log M_{Z(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})}(t) &= \log \mathbb{E}\exp\left(tZ(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})\right) \\ &= \gamma^2 t^2 \|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|^2 + \frac{1}{2}\sum_{k=3}^{\infty}\sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} \text{ for all } |t| < \frac{1}{8\gamma}(\sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})^2)^{-\frac{1}{2}}. \end{aligned} \tag{B.27}$$

Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be as in Lemma B.1 and Lemma B.2, respectively, and define $\mathcal{Q}_0 = \mathcal{Q}_1 \bigcap Q_2$. On $\mathcal{Q}_0$, there exists a constant $C > 0$ such that $\|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}\| < C$ (by (B.8)) and $\|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\| < C$ (since $\mathsf{H}_{\boldsymbol{\eta}}^{\circ} \in L^2(\mathcal{X}^2, P^2)$). Then by (B.26) and (B.27), both MGF's exists for $|t| \leqslant \frac{1}{8\gamma C}$ on $\mathcal{Q}_0$. Hereafter, we will assume that the we are on the set $\mathcal{Q}_0$. Using $\sum_{s=1}^{\infty} \lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})^2 = \|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|^2 < \infty$ gives, for all $s \geqslant 1$,

$$\left|\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})\right| \leqslant \frac{\|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|}{\sqrt{s}} \leqslant \frac{C}{\sqrt{s}}. \tag{B.28}$$

Then notice that for all $|t| \leqslant \frac{1}{8\gamma C} < \frac{1}{8\gamma}\|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|^{-1}$,

$$\sum_{k=3}^{\infty}\sum_{s=1}^{\infty}\left|\frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k}\right| \leqslant \sum_{k=3}^{\infty}\sum_{s=1}^{\infty} \frac{(2\gamma)^k\|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|^k}{(8\gamma)^k\|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|^k s^{k/2}k} \leqslant \sum_{k=3}^{\infty}\sum_{s=1}^{\infty} \frac{1}{4^k s^{3/2}k} < \infty, \tag{B.29}$$

and hence the second term in (B.27) is absolutely summable. Similarly, since $\sum_{s=1}^{\infty} \lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})^2 = \|\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ}\|^2 \leqslant C$,

$$\left|\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})\right| \leqslant \frac{C}{\sqrt{s}}, \tag{B.30}$$

for all $s \geqslant 1$, and the second term in (B.26) is also absolutely summable for all $|t| \leqslant \frac{1}{8\gamma C}$. Now, for any $N \geqslant 1$ and for all $|t| \leqslant \frac{1}{8\gamma C}$,

$$\begin{aligned}
&\left| \sum_{k=3}^{\infty} \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} - \sum_{k=3}^{\infty} \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} \right| \\
&\qquad \leqslant \left| \sum_{k=3}^{N} \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} - \sum_{k=3}^{N} \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} \right| \\
&\qquad\quad + \sum_{k=N+1}^{\infty} \left[\left| \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} - \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} \right| \right].
\end{aligned} \tag{B.31}$$

Note that the first term in (B.31) converges to zero as $m \to 0$ on $\mathcal{Q}_0$ (by (B.9)). Therefore, it suffices to show that the second term converges to zero in (B.31). Towards this, note that for $k \geqslant 3$, by (B.28) and (B.30),

$$\begin{aligned}
\left| \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} - \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} \right| &\leqslant \frac{(2\gamma)^k|t|^k}{k} \sum_{s=1}^{\infty} \left\{ \left|\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})\right|^k + \left|\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})\right|^k \right\} \\
&\leqslant \frac{2C^k(2\gamma)^k|t|^k}{k} \sum_{s=1}^{\infty} \frac{1}{s^{k/2}} \\
&\leqslant \frac{2C^k(2\gamma)^k|t|^k}{k} \sum_{s=1}^{\infty} \frac{1}{s^{3/2}}.
\end{aligned}$$

Then for $|t| \leqslant \frac{1}{8\gamma C}$,

$$\sum_{k=N+1}^{\infty} \left[\left| \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\hat{\mathbf{H}}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} - \sum_{s=1}^{\infty} \frac{(2\gamma\lambda_s(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})t)^k}{k} \right| \right] \leqslant 4 \sum_{s=1}^{\infty} \frac{1}{s^{3/2}} \sum_{k=N+1}^{\infty} \frac{1}{4^k k},$$

which converges to zero as $m \to \infty$ and then $N \to \infty$. This implies the RHS of (B.31) converges to zero as $m \to \infty$ and then $N \to \infty$. Thus, by (B.26), (B.27), and Lemma B.1, on the set $\mathcal{Q}_0$,

$$\lim_{m\to\infty} M_{Y_m|\mathscr{X}_m}(t) = M_{Z(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})}(t), \text{ for all } |t| < \frac{1}{8\gamma C}.$$

Hence, recalling (B.25), $Y_m|\mathscr{X}_m \xrightarrow{D} Z(\mathsf{H}_{\boldsymbol{\eta}}^{\circ})$ on the set $\mathcal{Q}_0$. Since $\mathbb{P}(\mathcal{Q}_0) = 1$, this completes the proof of Lemma B.3. □

Appendix C. Proof of Theorem 5.1

Suppose $\mathsf{H} \in L^2(\mathcal{X}^2, P^2)$ is a measurable and symmetric function (not necessarily positive definite) and recall the definition of $\mathrm{MMD}^2[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n]$ from (A.1). Note that

$$\mathrm{MMD}^2[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n] = \mathcal{W}_{\mathscr{X}_m} + \mathcal{W}_{\mathscr{Y}_n} - 2\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n} = \mathcal{W}^{\circ}_{\mathscr{X}_m} + \mathcal{W}^{\circ}_{\mathscr{Y}_n} - 2\mathcal{B}^{\circ}_{\mathscr{X}_m,\mathscr{Y}_n}, \tag{C.1}$$

where H° is as in (2.13) and

$$\mathcal{W}^\circ_{\mathscr{X}_m} := \frac{1}{m(m-1)} \sum_{1\leqslant i\neq j\leqslant m} \mathsf{H}^\circ(X_i, X_j) \text{ and } \mathcal{W}^\circ_{\mathscr{Y}_n} := \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} \mathsf{H}^\circ(Y_i, Y_j)$$

and

$$\mathcal{B}^\circ_{\mathscr{X}_m,\mathscr{Y}_n} := \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \mathsf{H}^\circ(X_i, Y_j).$$

Therefore, to obtain the limiting distribution of $\mathrm{MMD}^2[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n]$ we need to derive the joint distribution of $(\mathcal{W}^\circ_{\mathscr{X}_m}, \mathcal{W}^\circ_{\mathscr{Y}_n}, \mathcal{B}^\circ_{\mathscr{X}_m,\mathscr{Y}_n})$ under H_1. To this end, recall the definition of the Hilbert-Schmidt operator $\mathcal{H}_{\mathsf{H}^\circ}$ from (A.3). This operator has countably many eigenvalues $\{\lambda_s\}_{s\geqslant 1}$ with eigenvectors $\{\phi_s\}_{s\geqslant 1}$ satisfying:

$$\int_{\mathcal{X}} \mathsf{H}^\circ(x,y)\phi_s(y)\mathrm{d}P(y) = \lambda_s\phi_s(x) \text{ and } \int_{\mathcal{X}} \phi_s(x)\phi_{s'}(x)\mathrm{d}P(x) = \delta_{s,s'}, \tag{C.2}$$

for $s, s' \geqslant 1$ and $\delta_{s,s'} = 1$ if $s = s'$ and zero otherwise. Note that, since $\mathbb{E}_{X\sim P}[\mathsf{H}^\circ(X, y)] = 0$, for all $y \in \mathcal{X}$, whenever $\lambda_s \neq 0$, an application of Fubini's theorem and (C.2) implies that $\mathbb{E}_{X\sim P}[\phi_s(X)] = 0$. Moreover, (see, for example, [16, Theorem 4, Chapter X and Section XI.6] or [51, Theorem 8.94 and Theorem 8.83])

$$\sum_{s=1}^\infty \lambda_s^2 = \int_{\mathcal{X}^2} \mathsf{H}^\circ(x,y)^2\mathrm{d}x\mathrm{d}y = \|\mathsf{H}^\circ\|^2 < \infty, \tag{C.3}$$

and the spectral theorem,

$$\mathsf{H}^\circ(x,y) = \sum_{s=1}^\infty \lambda_s\phi_s(x)\phi_s(y), \tag{C.4}$$

where the convergence is in L^2.

The following result gives the joint distribution of $(\mathcal{W}^\circ_{\mathscr{X}_m}, \mathcal{W}^\circ_{\mathscr{Y}_n}, \mathcal{B}^\circ_{\mathscr{X}_m,\mathscr{Y}_n})$ and, hence, that of $\mathrm{MMD}^2[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n]$ under contiguous local alternatives (5.2) in the contamination model (5.1). The argument is similar to results in [12] on the limiting distribution of degenerate two-sample U-statistics for parametric contiguous alternatives.

Proposition C.1. *Suppose $\mathsf{H} \in L^2(\mathcal{X}^2, P^2)$ be a measurable and symmetric function. Then under H_1 as in (5.1) in the asymptotic regime (2.8),*

$$\begin{pmatrix} m\mathcal{W}^\circ_{\mathscr{X}_m} \\ n\mathcal{W}^\circ_{\mathscr{Y}_n} \\ \sqrt{mn}\mathcal{B}^\circ_{\mathscr{X}_m,\mathscr{Y}_n} \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \sum_{s=1}^\infty \lambda_s \left(W_s^2 - 1\right) \\ \sum_{s=1}^\infty \lambda_s \left(\left(W_s' + h\sqrt{1-\rho}L_s\right)^2 - 1\right) \\ \sum_{s=1}^\infty \lambda_s W_s \left(W_s' + h\sqrt{1-\rho}L_s\right) \end{pmatrix}, \tag{C.5}$$

where

- *$\{W_s, W_s' : s \geqslant 1\}$ are independent standard Gaussian random variables,*
- *$\{\lambda_s\}_{s\geqslant 1}$ are the eigenvalues (with repetitions) and the eigenvectors $\{\phi_s\}_{s\geqslant 1}$ of the Hilbert-Schimdt operator $\mathcal{H}_{\mathsf{H}^\circ_\eta}$ as in (C.2),*
- *$L_s := \mathbb{E}_{X\sim P}[\frac{\phi_s(X)g(X)}{f_P(X)}]$, for $s \geqslant 1$.*

Consequently, under H_1,

$$(m+n)\mathrm{MMD}^2[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{D} \tilde{Z}(\mathsf{H}) := \gamma \sum_{s=1}^\infty \lambda_s \left(\left(Z_s + \frac{h}{\sqrt{\gamma}}L_s\right)^2 - 1\right), \tag{C.6}$$

where $\gamma = \frac{1}{\rho(1-\rho)}$, $\{Z_s : s \geqslant 1\}$ *are i.i.d.* $\mathcal{N}(0,1)$. *Moreover,*

$$\mathbb{E}[\tilde{Z}(\mathsf{H})] = h^2 \sum_{s=1}^{\infty} \lambda_s L_s^2 = h^2 \mathbb{E}_{X,X' \sim P}\left[\mathsf{H}^{\circ}(X,X') \frac{g(X)g(X')}{f_P(X)f_P(X')}\right] < \infty \tag{C.7}$$

and the characteristic function of $\tilde{Z}(\mathsf{H})$ *at* $t \in \mathbb{R}$ *is given by:*

$$\Phi_{\tilde{Z}(\mathsf{H})}(t) := \mathbb{E}\left[e^{\iota t \tilde{Z}(\mathsf{H})}\right] = \frac{e^{\iota t h^2 \sum_{s=1}^{\infty} \lambda_s L_s^2 - \sum_{s=1}^{\infty}\left\{\iota\gamma\lambda_s t + \frac{\gamma h^2 \lambda_s^2 L_s^2 t^2}{(1-2\iota\lambda_s\gamma t)}\right\}}}{\prod_{s=1}^{\infty}\sqrt{1-2\iota\gamma\lambda_s t}}. \tag{C.8}$$

The proof of Proposition C.1 is given in Section C.1. Here, we show it can be used to complete the proof of Theorem 5.1. As in (A.5), for $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_r)^\top \in \mathbb{R}^r$,

$$\boldsymbol{\eta}^\top \mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] = \sum_{a=1}^{r} \eta_a \mathrm{MMD}^2[\mathsf{K}_a, \mathscr{X}_m, \mathscr{Y}_n] = \mathrm{MMD}^2[\mathsf{H}_{\boldsymbol{\eta}}, \mathscr{X}_m, \mathscr{Y}_n], \tag{C.9}$$

where $\mathsf{H}_{\boldsymbol{\eta}} := \sum_{a=1}^{r} \eta_a \mathsf{K}_a$. Then by Proposition C.1, under H_1,

$$\begin{aligned} Z_{m,n}(\mathsf{H}_{\boldsymbol{\eta}}) &:= (m+n)\mathrm{MMD}^2\left[\mathsf{H}_{\boldsymbol{\eta}}, \mathscr{X}_m, \mathscr{Y}_n\right] \\ &\xrightarrow{D} \tilde{Z}(\mathsf{H}_{\boldsymbol{\eta}}) = \gamma \sum_{s=1}^{\infty} \lambda_s \left(\left(Z_s + \frac{h}{\sqrt{\gamma}} L_s\right)^2 - 1\right), \end{aligned} \tag{C.10}$$

where, by (C.6), $\{\lambda_s\}_{s\geqslant 1}$ are the eigenvalues (with repetitions) and the eigenvectors $\{\phi_s\}_{s\geqslant 1}$ of the operator $\mathcal{H}_{\mathsf{H}_{\boldsymbol{\eta}}^{\circ}}$.

Note that by the linearity of the stochastic integral and arguments as in (3.7),

$$\sum_{a=1}^{r} \eta_a I_2(\mathsf{K}_a^{\circ}) = I_2(\mathsf{H}_{\boldsymbol{\eta}}^{\circ}) = \int_{\mathcal{X}}\int_{\mathcal{X}} \mathsf{H}_{\boldsymbol{\eta}}^{\circ}(x,y)\mathrm{d}\mathcal{Z}_P(x)\mathrm{d}\mathcal{Z}_P(y) \overset{D}{=} \sum_{s=1}^{\infty} \lambda_s \left(Z_s^2 - 1\right). \tag{C.11}$$

where $\{Z_s\}_{s\geqslant 1} \overset{D}{=} \left\{\int_{\mathcal{X}} \phi_s(x)\mathrm{d}\mathcal{Z}_P(x)\right\}_{s\geqslant 1}$. This also implies,

$$\begin{aligned} \sum_{s=1}^{\infty} \lambda_s \mathbb{E}_{X\sim P}\left[\frac{\phi_s(X)g(X)}{f_P(X)}\right] Z_s &= \sum_{s=1}^{\infty} \lambda_s Z_s \int_{\mathcal{X}} \phi_s(x)g(x)\mathrm{d}x \\ &\overset{D}{=} \sum_{s=1}^{\infty} \lambda_s \int \phi_s(x)g(x)\mathrm{d}x \left(\int_{\mathcal{X}} \phi_s(y)\mathrm{d}\mathcal{Z}_P(y)\right) \\ &= \left(\int_{\mathcal{X}} \mathsf{H}_{\boldsymbol{\eta}}^{\circ}(x,y)g(x)\mathrm{d}x\right)\mathrm{d}\mathcal{Z}_P(y) \\ &= I_1\left(\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\left[\frac{g}{f_P}\right]\right) = \sum_{a=1}^{r} \eta_a I_1\left(\mathsf{K}_a^{\circ}\left[\frac{g}{f_P}\right]\right), \end{aligned} \tag{C.12}$$

where the notations are as defined in Theorem 5.1. Also, by (C.31),

$$\begin{aligned} \sum_{s=1}^{\infty} \lambda_s L_s^2 &= \mathbb{E}_{X,X'\sim P}\left[\mathsf{H}_{\boldsymbol{\eta}}^{\circ}(X,X')\frac{g(X)g(X')}{f_P(X)f_P(X')}\right] \\ &= \sum_{a=1}^{r} \eta_a \mathbb{E}_{X,X'\sim P}\left[\mathsf{K}_a^{\circ}(X,X')\frac{g(X)g(X')}{f_P(X)f_P(X')}\right]. \end{aligned} \tag{C.13}$$

Using (C.11), (C.12) and (C.13) in (C.10) shows that $\tilde{Z}(\mathsf{H}_{\boldsymbol{\eta}}) \overset{D}{=} \boldsymbol{\eta}^\top G_{\mathcal{K},h}$ where $G_{\mathcal{K},h}$ is as defined in (5.3). This implies, from (C.9),

$$\boldsymbol{\eta}^\top \mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] \overset{D}{\to} \boldsymbol{\eta}^\top G_{\mathcal{K},h}.$$

Since $\boldsymbol{\eta} \in \mathbb{R}^r$ is arbitrary, this completes the proof of Theorem 5.1.

C.1. **Proof of Proposition C.1.** Fix $L \geqslant 1$ and define the L-truncated versions of $\mathcal{W}_{\mathscr{X}_m}$, $\mathcal{W}_{\mathscr{Y}_n}$, and $\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n}$ as follows:

$$\mathcal{W}^{\circ(L)}_{\mathscr{X}_m} := \frac{1}{m(m-1)} \sum_{s=1}^{L} \sum_{1\leqslant i\neq j\leqslant m} \lambda_s \phi_s(X_i)\phi_s(X_j) = \frac{1}{m(m-1)} \sum_{s=1}^{L} \lambda_s \left(\left(\sum_{i=1}^{m} \phi_s(X_i) \right)^2 - \sum_{i=1}^{m} \phi_s^2(X_i) \right),$$

$$\mathcal{W}^{\circ(L)}_{\mathscr{Y}_n} := \frac{1}{n(n-1)} \sum_{s=1}^{L} \sum_{1\leqslant i\neq j\leqslant n} \lambda_s \phi_s(Y_i)\phi_s(Y_j) = \frac{1}{n(n-1)} \sum_{s=1}^{L} \lambda_s \left(\left(\sum_{i=1}^{n} \phi_s(Y_i) \right)^2 - \sum_{i=1}^{n} \phi_s^2(Y_i) \right),$$

$$\mathcal{B}^{\circ(L)}_{\mathscr{X}_m,\mathscr{Y}_n} := \frac{1}{mn} \sum_{s=1}^{L} \lambda_s \sum_{i=1}^{m} \sum_{j=1}^{n} \phi_s(X_i)\phi_s(Y_j). \tag{C.14}$$

Define $U_{s,m} := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi_s(X_i)$ and $V_{s,n} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_s(Y_i)$, for $1 \leqslant s \leqslant L$ and the vectors

$$\boldsymbol{U}_m^{(L)} = (U_{s,m})_{1\leqslant s\leqslant L} \quad \text{and} \quad \boldsymbol{V}_n^{(L)} = (V_{s,n})_{1\leqslant s\leqslant L}. \tag{C.15}$$

Note that, under H_0, by the law of large numbers and (C.2)

$$\frac{1}{m} \sum_{i=1}^{m} \phi_s^2(X_i) \overset{P}{\to} \mathbb{E}_{X\sim P}[\phi_s(X)^2] = 1 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \phi_s^2(Y_i) \overset{P}{\to} \mathbb{E}_{Y\sim P}[\phi_s(Y)^2] = 1.$$

Hence, recalling (C.14), under H_0,

$$(m-1)\mathcal{W}^{\circ(L)}_{\mathscr{X}_m} = \sum_{s=1}^{L} \lambda_s \left(U_{s,m}^2 - 1\right) + o_P(1), \quad (n-1)\mathcal{W}^{\circ(L)}_{\mathscr{Y}_n} = \sum_{s=1}^{L} \lambda_s \left(V_{s,n}^2 - 1\right) + o_P(1), \tag{C.16}$$

and

$$\sqrt{mn}\mathcal{B}^{\circ(L)}_{\mathscr{X}_m,\mathscr{Y}_n} := \sum_{s=1}^{L} \lambda_s U_{s,m} V_{s,n}. \tag{C.17}$$

Therefore, obtain the joint distribution of $(\mathcal{W}^{\circ(L)}_{\mathscr{X}_m}, \mathcal{W}^{\circ(L)}_{\mathscr{Y}_n}, \mathcal{B}^{\circ(L)}_{\mathscr{X}_m,\mathscr{Y}_n})^\top$ it suffices to find the joint distribution of

$$\boldsymbol{Q}^{(L)}_{m,n} := \left(\sum_{s=1}^{L} \lambda_s \left(U_{s,m}^2 - 1\right), \sum_{s=1}^{L} \lambda_s \left(V_{s,n}^2 - 1\right), \sum_{s=1}^{L} \lambda_s U_{s,m} V_{s,n} \right)^\top. \tag{C.18}$$

This is derived in the following lemma. Here, $\mathbf{0}$ denotes the zero vector in $\mathbb{R}^L$ and $\boldsymbol{I}_{2L}$ denotes the $2L \times 2L$ identity matrix.

Lemma C.1. *Fix $L \geqslant 1$ and suppose $\boldsymbol{U}_m^{(L)}$ and $\boldsymbol{V}_n^{(L)}$ be as defined in (C.15). Then under H_1 as in (5.2) the following hold in the asymptotic regime (2.8),*

$$\begin{pmatrix} \boldsymbol{U}_m^{(L)} \\ \boldsymbol{V}_n^{(L)} \end{pmatrix} \overset{D}{\to} \mathcal{N}_{2L} \left(\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\theta} \end{pmatrix}, \boldsymbol{I}_{2L} \right). \tag{C.19}$$

where $\boldsymbol{\theta} := h\sqrt{1-\rho}\cdot(L_1, L_2 \ldots, L_r)^\top$ *and* L_s *as in Proposition C.1. Consequently,*

$$\begin{pmatrix} (m-1)\mathcal{W}^{\circ(L)}_{\mathscr{X}_m} \\ (n-1)\mathcal{W}^{\circ(L)}_{\mathscr{Y}_n} \\ \sqrt{mn}\mathcal{B}^{\circ(L)}_{\mathscr{X}_m,\mathscr{Y}_n} \end{pmatrix} \xrightarrow{D} \boldsymbol{Q}^{(L)} := \begin{pmatrix} \sum_{s=1}^{L}\lambda_s\left(W_s^2-1\right) \\ \sum_{s=1}^{L}\lambda_s\left(\left(W_s'+h\sqrt{1-\rho}L_s\right)^2-1\right) \\ \sum_{s=1}^{L}\lambda_s W_s\left(W_s'+h\sqrt{1-\rho}L_s\right) \end{pmatrix}, \tag{C.20}$$

where $\{W_s, W_s' : s \geqslant 1\}$ *are independent standard Gaussian random variables.*

Proof. To prove (C.19) we will first derive the joint distribution of $\boldsymbol{U}_m^{(L)}$, $\boldsymbol{V}_n^{(L)}$, and the log-likelihood ratio, and then invoke LeCam's third lemma [64, Example 6.7]. For the hypothesis in (5.1) the likelihood ratio L_N is given by:

$$L_N := \sum_{i=1}^{n}\log\left[\frac{\left(1-\frac{h}{\sqrt{N}}\right)f_P(Y_i)+\frac{h}{\sqrt{N}}g(Y_i)}{f_P(Y_i)}\right].$$

By local asymptotic normality (see, for example, [64, Chapter 7]), L_N can be written as:

$$L_N = \dot{L}_N - \frac{(1-\rho)h^2}{2}\cdot\delta_{f_P,g} + o_P(1),$$

where

$$\dot{L}_N = \frac{h}{\sqrt{N}}\sum_{i=1}^{n}\left(\frac{g(Y_i)}{f_P(Y_i)}-1\right) \text{ and } \delta_{f_P,g} := \int_{\mathcal{X}}\left(\frac{g(x)}{f_P(x)}-1\right)^2 f_P(x)\mathrm{d}x.$$

Note that $\mathrm{Cov}[U_{s,m}\dot{L}_N] = 0$ and

$$\begin{aligned}\mathrm{Cov}[V_{s,n}\dot{L}_N] = \mathbb{E}[V_{s,n}\dot{L}_N] &= \sqrt{\frac{n}{N}}\cdot\frac{h}{n}\sum_{i=1}^{n}\mathbb{E}\left[\phi_s(Y_i)\left(\frac{g(Y_i)}{f_P(Y_i)}-1\right)\right] \\ &\xrightarrow{P} h\sqrt{1-\rho}\cdot\mathbb{E}_{X\sim P}\left[\frac{\phi_s(X)g(X)}{f_P(X)}\right],\end{aligned}$$

by the law of large numbers and the fact $\mathbb{E}_{X\sim P}[\phi_s(X)] = 0$ (since we can assume $\lambda_s \neq 0$). Hence, by the multivariate central limit theorem, under H_0,

$$\begin{pmatrix} \boldsymbol{U}_m^{(L)} \\ \boldsymbol{V}_n^{(L)} \\ \dot{L}_N \end{pmatrix} \xrightarrow{D} \mathcal{N}_{2L+1}\left(\begin{pmatrix} \mathbf{0}_{L\times 1} \\ \mathbf{0}_{L\times 1} \\ -\frac{(1-\rho)h^2}{2}\delta_{f_P,g} \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_L & \mathbf{0}_{L\times L} & \mathbf{0}_{L\times 1} \\ \mathbf{0}_{L\times L} & \boldsymbol{I}_L & \boldsymbol{\theta} \\ \mathbf{0}_{1\times L} & \boldsymbol{\theta}^\top & (1-\rho)h^2\delta_{f_P,g} \end{pmatrix}\right) \tag{C.21}$$

where $\boldsymbol{\theta}$ is as defined in Lemma C.1 and $\mathbf{0}_{K\times L}$ is the $K\times L$ zero-matrix, for $K, L \geqslant 1$. Then by LeCam's third lemma [64, Example 6.7] the result in (C.19) follows.

Now, since $\boldsymbol{Q}_{m,n}^{(L)}$ (recall (C.18)) is a continuous function of $\boldsymbol{U}_m^{(L)}$ and $\boldsymbol{V}_n^{(L)}$, the result in (C.20) follows from (C.16), (C.17), (C.21), and the continuous mapping theorem. □

Next, we show that $\boldsymbol{Q}^{(L)}$ (as defined in Lemma C.1) converges as $L\to\infty$.

Lemma C.2. *Let* $\boldsymbol{Q}^{(L)}$ *be as defined in* (C.20)*. Then as* $L\to\infty$*,*

$$\boldsymbol{Q}^{(L)} \xrightarrow{L^2} \boldsymbol{Q} := \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \end{pmatrix} := \begin{pmatrix} \sum_{s=1}^{\infty}\lambda_s\left(W_s^2-1\right) \\ \sum_{s=1}^{\infty}\lambda_s\left(\left(W_s'+h\sqrt{1-\rho}L_s\right)^2-1\right) \\ \sum_{s=1}^{\infty}\lambda_s W_s\left(W_s'+h\sqrt{1-\rho}L_s\right), \end{pmatrix}, \tag{C.22}$$

where $\{W_s, W_s' : s \geqslant 1\}$ *are independent standard Gaussian random variables.*

Proof. Note that

$$\sum_{s=1}^{\infty} \mathrm{Var}\left[\lambda_s(W_s^2-1)\right] = 2\sum_{s=1}^{\infty}\lambda_s^2 < \infty,$$

by (C.3). Hence, as $L \to \infty$,

$$\sum_{s=1}^{L} \lambda_s\left(W_s^2-1\right) \stackrel{L^2}{\to} \sum_{s=1}^{\infty}\lambda_s\left(W_s^2-1\right) = Q_1.$$

Next, we denote $\theta_s := h\sqrt{1-\rho}L_s = h\sqrt{1-\rho}\mathbb{E}_{X\sim P}[\frac{\phi_s(X)g(X)}{f_P(X)}]$. Then by the Cauchy-Schwarz inequality,

$$\theta_s^2 \leqslant h^2(1-\rho)\mathbb{E}_{X\sim P}[\phi_s(X)^2]\mathbb{E}_{X\sim P}\left[\frac{g(X)^2}{f_P(X)^2}\right] = h^2(1-\rho)\mathbb{E}_{X\sim P}\left[\frac{g(X)^2}{f_P(X)^2}\right] < \infty, \tag{C.23}$$

since $\mathbb{E}_{X\sim P}\left[\phi_s(X)^2\right] = 1$, for all $s \geqslant 1$, and $\mathbb{E}_{X\sim P}[\frac{g(X)^2}{f_P(X)^2}] < \infty$ by Assumption 5.1. Then

$$\sum_{s=1}^{\infty}\mathrm{Var}\left[\lambda_s W_s\left(W_s'+\theta_s\right)\right] = \sum_{s=1}^{\infty}\lambda_s^2\mathbb{E}[W_s^2]\mathbb{E}\left[\left(W_s'+\theta_s\right)^2\right] = \sum_{s=1}^{\infty}\lambda_s^2(1+\theta_s^2) < \infty, \tag{C.24}$$

by (C.3) and (C.23). Hence, as $L \to \infty$,

$$\sum_{s=1}^{L}\lambda_s W_s(W_s'+\theta_s)^2 \stackrel{L^2}{\to} \sum_{s=1}^{\infty}\lambda_s W_s(W_s+\theta_s)^2 = Q_3.$$

It remains to establish the convergence to Q_2 in (C.22). Tp this end, fix $L \geqslant 1$ and denote

$$Q_2^{(L)} := \sum_{s=1}^{L}\lambda_s\left(\left(W_s'+\theta_s\right)^2-1\right).$$

Then for $L' > L \geqslant 1$,

$$\mathbb{E}\left[\left(Q_2^{(L')}-Q_2^{(L)}\right)^2\right] \leqslant \mathrm{Var}\left[\left(Q_2^{(L')}-Q_2^{(L)}\right)\right] + \left(\mathbb{E}\left[Q_2^{(L')}-Q_2^{(L)}\right]\right)^2. \tag{C.25}$$

This implies,

$$\begin{aligned}\mathrm{Var}\left[\left(Q_2^{(L')}-Q_2^{(L)}\right)\right] = \sum_{s=L+1}^{L'}\mathrm{Var}\left[\lambda_s\left(\left(W_s'+\theta_s\right)^2-1\right)\right] &= \sum_{s=L+1}^{L'}\lambda_s^2\,\mathrm{Var}\left[\left(W_s'+\theta_s\right)^2\right] \\ &= 2\sum_{s=L+1}^{L'}\lambda_s^2(1+2\theta_s^2) \to 0, \end{aligned} \tag{C.26}$$

as $L, L' \to \infty$, using (C.3) and (C.23). Next consider,

$$\begin{aligned}\left|\mathbb{E}\left[Q_2^{(L')}-Q_2^{(L)}\right]\right| = \sum_{s=L+1}^{L'}\lambda_s\theta_s^2 &= h\sqrt{1-\rho}\sum_{s=L+1}^{L'}\lambda_s\left(\int_{\mathcal{X}}\phi_s(x)g(x)\mathrm{d}x\right)^2 \\ &= h\sqrt{1-\rho}\int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{s=L+1}^{L'}\lambda_s\phi_s(x)\phi_s(y)g(y)g(y)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

Denote $C := h\sqrt{1-\rho}$ and $M := \mathbb{E}_{X\sim P}[\frac{g(X)^2}{f_P(X)^2}] < \infty$ (by Assumption 5.1). Then by the Cauchy-Schwarz inequality,

$$\left|\mathbb{E}\left[Q_2^{(L')} - Q_2^{(L)}\right]\right|^2 \leqslant C^2M^2 \int_{\mathcal{X}}\int_{\mathcal{X}} \left(\sum_{s=L+1}^{L'} \lambda_s \phi_s(x)\phi_s(y)\right)^2 f_P(x)f_P(y)\mathrm{d}x\mathrm{d}y \to 0, \tag{C.27}$$

as $L, L' \to \infty$, since the convergence in (C.4) is in L^2. Combining (C.27) and (C.26) with (C.25) it follows that $Q_2^{(L)}$ converges in L^2 to Q_2. This completes the proof of Lemma C.2. □

Next, we show that $(\mathcal{W}^{\circ(L)}_{\mathscr{X}_m}, \mathcal{W}^{\circ(L)}_{\mathscr{Y}_n}, \mathcal{B}^{\circ(L)}_{\mathscr{X}_m,\mathscr{Y}_n})^\top$ (recall (C.14)) and $(\mathcal{W}^{\circ}_{\mathscr{X}_m}, \mathcal{W}^{\circ}_{\mathscr{Y}_n}, \mathcal{B}^{\circ}_{\mathscr{X}_m,\mathscr{Y}_n})^\top$ are asymptotically close.

Lemma C.3. *As $L \to \infty$,*

$$\sup_{m,n\geqslant 1} \mathbb{E}\left\| \begin{pmatrix} (m-1)\mathcal{W}^{\circ(L)}_{\mathscr{X}_m} \\ (n-1)\mathcal{W}^{\circ(L)}_{\mathscr{Y}_n} \\ \sqrt{mn}\mathcal{B}^{\circ(L)}_{\mathscr{X}_m,\mathscr{Y}_n} \end{pmatrix} - \begin{pmatrix} (m-1)\mathcal{W}^{\circ}_{\mathscr{X}_m} \\ (n-1)\mathcal{W}^{\circ}_{\mathscr{Y}_n} \\ \sqrt{mn}\mathcal{B}^{\circ}_{\mathscr{X}_m,\mathscr{Y}_n} \end{pmatrix} \right\|^2 \to 0.$$

Proof. Note that by (C.4) and Fubini's theorem,

$$\begin{aligned} \begin{pmatrix} (m-1)\mathcal{W}^{\circ}_{\mathscr{X}_m} \\ (n-1)\mathcal{W}^{\circ}_{\mathscr{Y}_n} \\ \sqrt{mn}\mathcal{B}^{\circ}_{\mathscr{X}_m,\mathscr{Y}_n} \end{pmatrix} &= \begin{pmatrix} \frac{1}{m}\sum_{1\leqslant i\neq j\leqslant m} \mathsf{H}^\circ(X_i, X_j) \\ \frac{1}{n}\sum_{1\leqslant i\neq j\leqslant n} \mathsf{H}^\circ(Y_i, Y_j) \\ \frac{1}{\sqrt{mn}}\sum_{i=1}^m\sum_{i=1}^n \mathsf{H}^\circ(X_i, Y_j) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{m}\sum_{s=1}^\infty \lambda_s \sum_{1\leqslant i\neq j\leqslant m} \phi_s(X_i)\phi_s(X_j) \\ \frac{1}{n}\sum_{s=1}^\infty \lambda_s \sum_{1\leqslant i\neq j\leqslant m} \phi_s(Y_i)\phi_s(Y_j) \\ \frac{1}{\sqrt{mn}}\sum_{s=1}^\infty \lambda_s \sum_{i=1}^m\sum_{j=1}^n \phi_s(X_i)\phi_s(Y_j) \end{pmatrix} \end{aligned} \tag{C.28}$$

where the existence of such infinite sums will be proved in the following. Using $\mathbb{E}_{X\sim P}[\phi_s(X)] = 0$ for $s \geqslant 1$, it is easy to show that for $s \neq s'$,

$$\mathbb{E}\left[\left(\sum_{1\leqslant i\neq j\leqslant m} \phi_s(X_i)\phi_s(X_j)\right)\left(\sum_{1\leqslant i\neq j\leqslant m} \phi_{s'}(X_i)\phi_{s'}(X_j)\right)\right] = 0$$

and using (C.2),

$$\mathbb{E}\left[\left(\sum_{1\leqslant i\neq j\leqslant m} \phi_s(X_i)\phi_s(X_j)\right)^2\right] = 4\mathbb{E}\left[\sum_{1\leqslant i<j\leqslant m} (\phi_s(X_i)\phi_s(X_j))^2\right] = 2m(m-1).$$

Define $c_m = \sqrt{2m(m-1)}$. Then $\{\frac{1}{c_m}\sum_{1\leqslant i\neq j\leqslant m} \phi_s(X_i)\phi_s(X_j)\}_{s\geqslant 1}$ is a collection of orthonormal random variables. Hence, by [38, Lemma 6.8] the following infinite sum

$$\sum_{s=1}^\infty \lambda_s \left(\frac{1}{c_m}\sum_{1\leqslant i\neq j\leqslant m} \phi_s(X_i)\phi_s(X_j)\right)$$

exists, which also proves the existence of the infinite sum in (C.28). Moreover,

$$\begin{aligned} \mathbb{E}\left[(m-1)\left(\mathcal{W}^{\circ}_{\mathscr{X}_m} - \mathcal{W}^{\circ(L)}_{\mathscr{X}_m}\right)\right]^2 &\leqslant \frac{m^2}{c_m^2}\mathbb{E}\left[\sum_{s=L+1}^\infty \lambda_s \left(\frac{1}{c_m}\sum_{1\leqslant i\neq j\leqslant n} \phi_s(X_i)\phi_s(X_j)\right)\right]^2 \\ &\leqslant \sum_{s=L+1}^\infty \lambda_s^2 \to 0, \end{aligned}$$

as $L \to \infty$ (recall (C.3)), uniformly in m, n. Similarly, it can be shown that

$$(n-1)^2 \mathbb{E}\left[\mathcal{W}_{\mathscr{Y}_n} - \mathcal{W}^{\circ(L)}_{\mathscr{Y}_n}\right]^2 \to 0 \text{ and } mn\mathbb{E}\left[\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n} - \mathcal{B}^{(L)}_{\mathscr{X}_m,\mathscr{Y}_n}\right]^2 \to 0$$

as $L \to \infty$, uniformly in m, n. □

Combining Lemmas C.1, C.2, C.3 and using [32, Lemma 6] we get,

$$\begin{pmatrix} (m-1)\mathcal{W}^{\circ}_{\mathscr{X}_m} \\ (n-1)\mathcal{W}^{\circ}_{\mathscr{Y}_n} \\ \sqrt{mn}\mathcal{B}^{\circ}_{\mathscr{X}_m,\mathscr{Y}_n} \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \sum_{s=1}^{\infty} \lambda_s \left(W_s^2 - 1\right) \\ \sum_{s=1}^{\infty} \lambda_s \left((W'_s + \theta_s)^2 - 1\right) \\ \sum_{s=1}^{\infty} \lambda_s W_s \left(W'_s + \theta_s\right) \end{pmatrix},$$

where $\theta_s := h\sqrt{1-\rho} \cdot \mathbb{E}_{X\sim P}[\frac{\phi_s(X)g(X)}{f_P(X)}]$, for $s \geqslant 1$. This establishes (C.5). Then (C.1) and the continuous mapping theorem gives,

$$\begin{aligned}
&(m+n)\mathrm{MMD}^2\left[\mathsf{H}, \mathscr{X}_m, \mathscr{Y}_n\right] \\
&= (m+n)\mathcal{W}^{\circ}_{\mathscr{X}_m} + (m+n)\mathcal{W}^{\circ}_{\mathscr{Y}_n} - 2(m+n)\mathcal{B}^{\circ}_{\mathscr{X}_m,\mathscr{Y}_n} \\
&\xrightarrow{D} \frac{1}{\rho}\sum_{s=1}^{\infty} \lambda_s \left(W_s^2 - 1\right) + \frac{1}{1-\rho}\sum_{s=1}^{\infty} \lambda_s \left(\left(W'_s + \theta_s\right)^2 - 1\right) + \frac{2}{\sqrt{\rho(1-\rho)}}\sum_{s=1}^{\infty} \lambda_s W_s \left(W'_s + \theta_s\right) \\
&= \sum_{s=1}^{\infty} \lambda_s \left(\left(\frac{1}{\sqrt{\rho}}W_s - \frac{1}{\sqrt{1-\rho}}W'_s + h\mathbb{E}_{X\sim P}\left[\frac{\phi_s(X)g(X)}{f_P(X)}\right]\right)^2 - \frac{1}{\rho(1-\rho)}\right) \qquad \text{(C.29)} \\
&\overset{D}{=} \frac{1}{\rho(1-\rho)}\sum_{s=1}^{\infty} \lambda_s \left(\left(Z_s + h\sqrt{\rho(1-\rho)}\mathbb{E}_{X\sim P}\left[\frac{\phi_s(X)g(X)}{f_P(X)}\right]\right)^2 - 1\right),
\end{aligned}$$

where $\{Z_s\}_{s\geqslant 1}$ are i.i.d. $\mathcal{N}(0,1)$ and the rearrangement of the terms in (C.29) can be justified by truncation and taking limits. This completes the proof of (C.6).

To show (C.7) note that

$$\begin{aligned}
\mathbb{E}[\tilde{Z}(\mathsf{H})] = h^2 \sum_{s=1}^{\infty} \lambda_s L_s^2 &= h^2 \sum_{s=1}^{\infty} \lambda_s \left(\int_{\mathcal{X}} \phi_s(x)g(x)\mathrm{d}x\right)^2 \\
&= h^2 \int_{\mathcal{X}}\int_{\mathcal{X}} \sum_{s=1}^{\infty} \lambda_s \phi_s(x)\phi_s(y)g(y)g(y)\mathrm{d}x\mathrm{d}y \qquad \text{(C.30)} \\
&= h^2 \int_{\mathcal{X}}\int_{\mathcal{X}} \mathsf{H}^{\circ}(x,y)g(y)g(y)\mathrm{d}x\mathrm{d}y \qquad \text{(by (C.3))} \\
&= h^2 \mathbb{E}_{X,X'\sim P}\left[\mathsf{H}^{\circ}(X,X')\frac{g(X)g(X')}{f_P(X)f_P(X')}\right] < \infty, \qquad \text{(C.31)}
\end{aligned}$$

where the exchange of expectation and integral in (C.30) is valid by arguments similar to (C.27) and finiteness of the expectation follows by the Cauchy-Schwarz inequality, Assumption 5.1, and the fact $\mathsf{H}^{\circ} \in L^2(\mathcal{X}^2, P^2)$. Finally, the expression for the characteristic function in (C.8) follows from [31, Theorem 6.2], since

$$\sum_{s=1}^{\infty} \lambda_s^2 L_s^2 = \sum_{s=1}^{\infty} \lambda_s^2 \left(\mathbb{E}_{X\sim P}\left[\frac{\phi_s(X)g(X)}{f_P(X)}\right]\right)^2 < \infty,$$

by arguments as in (C.24). This completes the proof of Proposition C.1. □

Appendix D. Proof of Theorem 6.1

For $1 \leqslant a \leqslant r$, let

$$\tilde{\Delta}_a^{(1)}(x) := \mathbb{E}_{X' \sim P, Y, Y' \sim Q}[h_a(x, X', Y, Y')] - \mathrm{MMD}^2[\mathcal{F}_a, P, Q] \tag{D.1}$$

and

$$\tilde{\Delta}_a^{(1)}(y) = \mathbb{E}_{X, X' \sim P, Y' \sim Q}[h_a(X, X', y, Y')] - \mathrm{MMD}^2[\mathcal{F}_a, P, Q]. \tag{D.2}$$

Recalling (6.1), the first-order Hoeffding's projection for $\mathrm{MMD}^2\left[\mathcal{K}, \mathscr{X}_m, sY_n\right] - \mathrm{MMD}^2[\mathcal{F}, P, Q]$ is given by,

$$\hat{\boldsymbol{U}}_{m,n} = \frac{2}{m} \sum_{i=1}^{m} \tilde{\boldsymbol{\Delta}}^{(1)}(X_i) + \frac{2}{n} \sum_{j=1}^{n} \tilde{\boldsymbol{\Delta}}^{(2)}(Y_j),$$

where $\tilde{\boldsymbol{\Delta}}^{(1)}(x) = (\tilde{\Delta}_a^{(1)}(x))_{1 \leqslant a \leqslant r}$ and $\tilde{\boldsymbol{\Delta}}^{(2)}(x) = (\tilde{\Delta}_a^{(2)}(x))_{1 \leqslant a \leqslant r}$. Then by [64, Theorem 12.6],

$$\left\| \sqrt{m+n} \left(\mathrm{MMD}^2\left[\mathsf{K}, P, Q\right] - \mathrm{MMD}^2[\mathcal{F}, P, Q] - \hat{\boldsymbol{U}}_{m,n} \right) \right\| = o_P(1). \tag{D.3}$$

By the multivariate central limit theorem,

$$\sqrt{m+n}\, \hat{\boldsymbol{U}}_{m,n} \xrightarrow{D} \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Gamma}),$$

where

$$\boldsymbol{\Gamma} = 4 \left(\rho \operatorname{Var}_{X \sim P}\left[\tilde{\boldsymbol{\Delta}}^{(1)}(X) \right] + (1-\rho) \operatorname{Var}_{Y \sim Q}\left[\tilde{\boldsymbol{\Delta}}^{(2)}(Y) \right] \right) = \boldsymbol{\Sigma},$$

since, recalling (6.3) and (D.1), $\operatorname{Var}_{X \sim P}\left[\tilde{\boldsymbol{\Delta}}^{(1)}(X)\right] = \operatorname{Var}_{X \sim P}[\boldsymbol{\Delta}^{(1)}(X)]$ and, from (6.4) and (D.2), $\operatorname{Var}_{Y \sim Q}\left[\tilde{\boldsymbol{\Delta}}^{(2)}(Y)\right] = \operatorname{Var}_{Y \sim Q}[\boldsymbol{\Delta}^{(2)}(Y)]$. This together with (D.3) completes the proof of Theorem 6.1. □

Appendix E. Invertibility of Kernel Matrices

In this section we discuss the invertibility of matrix $\boldsymbol{\Sigma}_{H_0}$ (recall the definition from (2.12)). Throughout this section we will assume that the underlying space $\mathcal{X} = \mathbb{R}^d$ and the distribution P satisfy the following:

Assumption E.1. *Suppose $\mathcal{X} = \mathbb{R}^d$ and the distribution P has a density with respect to the Lebesgue measure on $\mathbb{R}^d$ with full support.*

The following proposition gives a set of general conditions under which $\boldsymbol{\Sigma}_{H_0}$ is non-singular. In Corollary E.1 we will show that these conditions are satisfied by the commonly used kernels, such as the Gaussian and Laplace kernels.

Proposition E.1. *Suppose Assumption E.1 holds and $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ be a collection of r distinct characteristic kernels such that:*

- *For every $x, y \in \mathbb{R}^d$ and $1 \leqslant a \leqslant r$,*

$$\lim_{\|z\| \to \infty} \mathsf{K}_a(x, z) = 0 \text{ and } \lim_{\|z\| \to \infty} \mathsf{K}_a(z, y) = 0. \tag{E.1}$$

- *For every collection $\{\alpha_a : 1 \leqslant a \leqslant r\}$ there exists a set $\Gamma \in \mathbb{R}^{2d}$ with $\mu(\Gamma) > 0$ such that*

$$\sum_{a=1}^{r} \alpha_a \mathsf{K}_a \neq 0 \quad \text{for all } (x, y) \in \Gamma. \tag{E.2}$$

Then $\boldsymbol{\Sigma}_{H_0}$ is non-singular.

Proof. Throughout the proof we will use μ to denote the Lebesgue measure in appropriate dimensions. Recall from (2.12) that $\mathbf{\Sigma}_{H_0} = ((\sigma_{ab}))_{1\leqslant a,b\leqslant r}$, where

$$\sigma_{ab} = \frac{2}{\rho^2(1-\rho)^2}\mathbb{E}\left[\mathsf{K}_a^\circ(X,X')\mathsf{K}_b^\circ(X,X')\right] = \frac{2}{\rho^2(1-\rho)^2}\operatorname{Cov}[\mathsf{K}_a^\circ(X,X'),\mathsf{K}_b^\circ(X,X')],$$

for $X, X' \sim P$. Hence, $\mathbf{\Sigma}_{H_0}$ is singular if and only if there exists $\alpha_1, \alpha_2, \ldots, \alpha_r \in \mathbb{R}^r$ such that

$$\sum_{a=1}^r \alpha_a \mathsf{K}_a^\circ(X,X') = 0 \quad \text{almost surely } P^2.$$

Then by Assumption E.1, there exists a set $A \subseteq \mathbb{R}^{2d}$ with $\mu(A^c) = 0$, such that

$$\sum_{a=1}^r \alpha_a \mathsf{K}_a^\circ(x,y) = 0, \tag{E.3}$$

for all $(x,y) \in A$. Then considering $h(x,y) = \sum_{a=1}^r \alpha_a \mathsf{K}_a(x,y)$ gives, for $(x,y) \in A$,

$$\begin{aligned} h(x,y) &= \sum_{a=1}^r \alpha_a \left(\mathbb{E}[\mathsf{K}_a(x,X')] + \mathbb{E}[\mathsf{K}_a(X,y)] + \mathsf{K}_a^\circ(x,y) - \mathbb{E}[\mathsf{K}_a(X,X')]\right) \\ &= \sum_{a=1}^r \alpha_a \left(\mathbb{E}[\mathsf{K}_a(x,X')] + \mathbb{E}[\mathsf{K}_a(X,y)] - \mathbb{E}[\mathsf{K}_a(X,X')]\right) \qquad \text{(by (E.3))} \\ &= \sum_{a=1}^r \alpha_a \left(f_a(x) + g_a(y)\right) + L, \end{aligned} \tag{E.4}$$

where $L := -\sum_{a=1}^r \alpha_a \mathbb{E}[\mathsf{K}_a(X,X')]$, $f_a(x) := \mathbb{E}[\mathsf{K}_a(x,X')]$, and $g_a(y) := \mathbb{E}[\mathsf{K}_a(X,y)]$, for $1 \leqslant a \leqslant r$. For any $x \in \mathbb{R}^d$ consider,

$$A_x := \left\{y \in \mathbb{R}^d : (x,y) \in A\right\}.$$

Denote $B := \{x \in \mathbb{R}^d : \mu(A_x^c) = 0\}$. Then by Fubini's Theorem, $\mu(B) = 1$. Now, consider $x' \in B$ and a sequence $\{x_N\}_{N\geqslant 1}$ in B such that $\|x_N\| \to \infty$. Define,

$$A_\infty = A_{x'} \bigcap \left(\bigcap_{n\geqslant 1} A_{x_N}\right).$$

By definition $\mu(A_\infty^c) = 0$. Also, if $y \in A_\infty$, then $(x_N, y) \in A$ and recalling (E.4) we have,

$$h(x_N,y) = \sum_{a=1}^r \alpha_a f_a(x_N) + \sum_{a=1}^r \alpha_a g_a(y) + L, \tag{E.5}$$

for all $N \geqslant 1$. Similarly, we also have $h(x',y) = \sum_{a=1}^r \alpha_a f_a(x') + \sum_{a=1}^r \alpha_a g_a(y) + L$. This implies,

$$\sum_{a=1}^r \alpha_a g_a(y) = h(x',y) - \sum_{a=1}^r \alpha_a f_a(x') - L = h(x',y) + L_{x'}, \tag{E.6}$$

where $L_{x'} := -\sum_{a=1}^r \alpha_a f_a(x') - L$ is a constant depending on x'. Next, fixing $y' \in A_\infty$ we get,

$$\sum_{a=1}^r \alpha_a f_a(x_N) = h(x_N,y') - \sum_{a=1}^r b_a g_a(y') - L = h(x_N,y') + L_{y'}, \tag{E.7}$$

where $L_{y'} := -\sum_{a=1}^r b_a g_a(y') - L$ is a constant depending on y'. Thus, combining (E.5), (E.6), and (E.7),

$$h(x_N,y) = h(x',y) + h(x_N,y') + L_{x'} + L_{y'} + L, \tag{E.8}$$

for all $y \in A_\infty$. Now, since $\mu(A_\infty) = 1$ we can choose a sequence $\{y_M\}_{M\geqslant 1}$ in A_∞ such that $\|y_M\| \to \infty$. Then observe that,

$$h(x_N, y_M) = h(x', y_M) + h(x_N, y') + L_{x'} + L_{y'} + L.$$

Taking limits as $M \to \infty$ and then $N \to \infty$ and using condition (E.1) it follows that $L_{x'}+L_{y'}+L = 0$. Thus, from (E.8), for all $y \in A_\infty$, $h(x_N, y) = h(x', y) + h(x_N, y')$. Therefore, taking $N \to \infty$ and using (E.1) gives,

$$h(x', y) = \sum_{a=1}^{r} \alpha_a \mathsf{K}_a(x', y) = 0 \quad \text{for all } y \in A_\infty.$$

This implies, since x' is arbitrarily chosen from B and $\mu(B) = 1$,

$$\mu\left\{x \in \mathbb{R}^d : \mu\left\{y \in \mathbb{R}^d : \sum_{a=1}^{r} \alpha_a \mathsf{K}_a(x, y) \neq 0\right\} > 0\right\} = 0.$$

Therefore, by Fubini's theorem,

$$\mu\left\{(x, y) \in \mathbb{R}^{2d} : \sum_{a=1}^{r} \alpha_a \mathsf{K}_a(x, y) \neq 0\right\} = 0,$$

which contradicts (E.2). Thus, $\mathbf{\Sigma}_{H_0}$ is non-singular whenever (E.1) and (E.2) hold. □

Proposition E.1 gives general conditions under which the matrix $\mathbf{\Sigma}_{H_0}$ is invertible. Clearly, condition (E.1) is satisfied by most kernels. We will now show that condition (E.2) holds when $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ is a collection of r distinct Gaussian or Laplace kernels.

Corollary E.1. *Suppose Assumption E.1 holds and $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ is a collection of r distinct Gaussian or Laplace kernels, that is, K_a is either a Gaussian kernel or a Laplace kernel with bandwidth σ_a, for $1 \leqslant a \leqslant r$, with $\sigma_1 \neq \sigma_2 \neq \cdots \neq \sigma_r > 0$. Then the conditions of Proposition E.1 hold and, consequently, $\mathbf{\Sigma}_{H_0}$ is non-singular.*

Proof. Clearly, condition (E.1) holds for the Gaussian and the Laplace kernels. To show (E.2) assume for contradiction that there exists $\alpha_1, \alpha_2, \ldots, \alpha_r$ such that

$$\mu\left\{(x, y) \in \mathbb{R}^{2d} : \sum_{a=1}^{r} \alpha_a \mathsf{K}_a(x, y) \neq 0\right\} = 0.$$

Note that without loss of generality we can assume that $\alpha_a \neq 0, 1 \leqslant a \leqslant r$. Denote $D := \{x \in \mathbb{R}^d : \mu\{y \in \mathbb{R}^d : \sum_{a=1}^{r} \alpha_a \mathsf{K}_a \neq 0\} = 0\}$. By Fubini's theorem $\mu(D) = 1$. For any $x \in D$ denote,

$$D_x := \left\{y \in \mathbb{R}^d : \sum_{a=1}^{r} \alpha_a \mathsf{K}_a = 0\right\}.$$

By definition, $\mu(D_x) = 1$. Then we can find a sequence $\{y_M\}_{M\geqslant 1}$ in D_x such that $\|y_M\| \to \infty$ such that

$$\sum_{a=1}^{r} \alpha_a \mathsf{K}_a(x, y_M) = 0.$$

Then,

$$\sum_{a=2}^{r} \alpha_a \frac{\mathsf{K}_a(x, y_M)}{\mathsf{K}_1(x, y_M)} = -\alpha_1. \tag{E.9}$$

Now, we have the following cases:

- K_a *is a Gaussian kernel for all* $1 \leqslant a \leqslant r$: Without loss of generality, suppose $\sigma_1 = \arg\max_{1\leqslant a\leqslant r} \sigma_a$. Then by the definition of Gaussian kernel, it follows that,

$$\sum_{a=2}^{r} \alpha_a \frac{\mathsf{K}_a(x, y_M)}{\mathsf{K}_1(x, y_M)} \to 0,$$

  as $M \to \infty$. This contradicts (E.9).
- K_a *is a Laplace kernel for some* $1 \leqslant a \leqslant r$: Without loss of generality, suppose σ_1 be the largest bandwidth among the Laplace kernels. Then by the definition of Gaussian and Laplace kernels it follows that

$$\sum_{a=2}^{r} \alpha_a \frac{\mathsf{K}_a(x, y_M)}{\mathsf{K}_1(x, y_M)} \to 0,$$

  as $M \to \infty$. As in the previous case, this contradicts (E.9).

This completes the proof of Corollary E.1. □

Appendix F. Technical Lemmas

In this section we collect the proofs of various technical lemmas. We begin by showing the continuity of the characteristic function of the limiting distribution.

Lemma F.1. *Let* $\Phi(\boldsymbol{\eta})$ *be as in* (3.4). *Then* $\Phi(\mathbf{0}) = 1$ *and* $\Phi(\boldsymbol{\eta})$ *is continuous at* $\mathbf{0} \in \mathbb{R}^r$.

Proof. Note that when $\boldsymbol{\eta} = \mathbf{0}$, the only eigenvalue of the operator $\mathcal{H}_{\mathcal{K},\eta}$ is zero and, hence, $\Phi(\mathbf{0}) = 1$.

For showing continuity at $\boldsymbol{\eta} = \mathbf{0}$, recall that for all $\boldsymbol{\eta} \in \mathbb{R}^r$, $\sum_{\lambda\in\Lambda(\boldsymbol{\eta})} \lambda^2 < \infty$, since $\mathcal{H}_{\mathcal{K},\boldsymbol{\eta}}$ is a Hilbert-Schmidt operator. Then by Fubini's theorem,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{\lambda\in\Lambda(\boldsymbol{\eta})} \lambda(Z_\lambda^2 - 1)\right)^2\right] &= \mathbb{E}\left[\sum_{\lambda} \lambda^2(Z_\lambda^2 - 1) + \sum_{\lambda_1\neq\lambda_2} \lambda_1\lambda_2 \left(Z_{\lambda_1}^2 - 1\right)\left(Z_{\lambda_2}^2 - 1\right)\right] \\
&= \sum_{\lambda} \lambda^2 \mathbb{E}(Z_\lambda^2 - 1)^2 \\
&= 2\sum_{\lambda} \lambda^2. \qquad \text{(F.1)}
\end{aligned}$$

By the spectral theorem (see [51, Theorem 6.35]) and recalling (A.8),

$$\sum_{\lambda\in\Lambda(\boldsymbol{\eta})} \lambda^2 = \left\|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\right\|^2.$$

Clearly, $\lim_{\boldsymbol{\eta}\to\mathbf{0}} \|\mathsf{H}_{\boldsymbol{\eta}}^{\circ}\|^2 = 0$ and thus by (A.6) and (F.1) and we conclude, $Z(\mathsf{H}_{\boldsymbol{\eta}}) \xrightarrow{L^2} 0$, as $\boldsymbol{\eta} \to \mathbf{0}$. Hence, by (A.9) and the Dominated Convergence Theorem,

$$\lim_{\boldsymbol{\eta}\to 0} \Phi(\boldsymbol{\eta}) = \lim_{\boldsymbol{\eta}\to\mathbf{0}} \mathbb{E}\left[e^{\iota Z(\mathsf{H}_{\boldsymbol{\eta}})}\right] = 1.$$

This shows the continuity of $\Phi(\boldsymbol{\eta})$ at $\boldsymbol{\eta} = \mathbf{0} \in \mathbb{R}^r$. □

Next, we compute the MGF of the random variable $Z(\mathsf{H})$ as defined (A.2).

Lemma F.2. *The MGF of $Z(\mathsf{H})$, as defined in (A.2), exists for all $|t| < \frac{\rho(1-\rho)}{8}\|\mathsf{H}^\circ\|^{-1}$ and is given by,*

$$\log M_{Z(\mathsf{H})}(t) := \log \mathbb{E}\left[e^{tZ(\mathsf{H})}\right] = \left(\frac{t}{\rho(1-\rho)}\right)^2 \|\mathsf{H}^\circ\|^2 + \frac{1}{2}\sum_{K=3}^{\infty}\sum_{s=1}^{\infty}\frac{\left(\frac{2}{\rho(1-\rho)}\lambda_s t\right)^K}{K}$$

where $\{\lambda_s : s \geqslant 1\}$ are the eigenvalues of the operator $\mathcal{H}_{\mathsf{H}^\circ}$.

Proof. Define $\gamma = \frac{1}{\rho(1-\rho)}$. Since $\sum_{s=1}^{\infty}\lambda_s^2 < \infty$, by [8, Proposition 7.1] we have,

$$M_{Z(\mathsf{H})}(t) := \mathbb{E}\left[e^{tZ(\mathsf{H})}\right] = \prod_{s=1}^{\infty}\frac{e^{-\gamma\lambda_s t}}{\sqrt{1-2\gamma\lambda t}}, \tag{F.2}$$

for all $|t| < \frac{1}{8\gamma}(\sum_{s=1}^{\infty}\lambda_s^2)^{-\frac{1}{2}} = \frac{1}{8\gamma}\|\mathsf{H}^\circ\|^{-1}$ (since $\|\mathsf{H}^\circ\|^2 = \sum_{s=1}^{\infty}\lambda_s^2$). Fix t such that $|t| < \frac{1}{8\gamma}\|\mathsf{H}^\circ\|^{-1}$, then taking log on both sides of (F.2) and expanding we get,

$$\log M_{Z(\mathsf{H})}(t) = \sum_{s=1}^{\infty}\left\{-\gamma\lambda_s t + \frac{1}{2}\left(\sum_{k=1}^{\infty}\frac{(2\gamma\lambda_s t)^k}{k}\right)\right\} = \frac{1}{2}\sum_{s=1}^{\infty}\sum_{k=2}^{\infty}\frac{(2\gamma\lambda_s t)^k}{k}. \tag{F.3}$$

Using $\|\mathsf{H}^\circ\|^2 = \sum_{s=1}^{\infty}\lambda_s^2$ and (F.3) we get,

$$\log M_{Z(\mathsf{H})}(t) = \gamma^2\|\mathsf{H}^\circ\|^2 t^2 + \frac{1}{2}\sum_{s=1}^{\infty}\sum_{k=3}^{\infty}\frac{(2\gamma\lambda_s t)^k}{k}. \tag{F.4}$$

Using the bounds $|t| < \frac{1}{8\gamma}\|\mathsf{H}^\circ\|^{-1}$ and $|\lambda_s| \leqslant \frac{\|\mathsf{H}^\circ\|}{\sqrt{s}}$ for all $s \in \mathbb{N}$, where $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots$ (again using $\|\mathsf{H}^\circ\|^2 = \sum_{s=1}^{\infty}|\lambda_s|^2$) gives,

$$\sum_{s=1}^{\infty}\sum_{k=3}^{\infty}\frac{|2\gamma\lambda_s t|^k}{k} \leqslant \sum_{s=1}^{\infty}\sum_{k=3}^{\infty}\frac{|\lambda_s|^k}{4^k\|\mathsf{H}^\circ\|^k k} \leqslant \sum_{s=1}^{\infty}\sum_{k=3}^{\infty}\frac{1}{s^{\frac{3}{2}}4^k k} < \infty,$$

Therefore, by Fubini's Theorem we can interchange the order of the sum in (F.4) to get,

$$\log M_{Z(\mathsf{H})}(t) = \gamma^2\|\mathsf{H}^\circ\|^2 t^2 + \frac{1}{2}\sum_{k=3}^{\infty}\sum_{s=1}^{\infty}\frac{(2\gamma\lambda_s t)^k}{k},$$

for all $|t| < \frac{1}{8\gamma}\|\mathsf{H}^\circ\|^{-1}$, which completes the proof. □

In the next lemma we show that the row sums of a characteristic kernel is asymptotically close to its expected value in an L_2 sense. This is used in the proof of Proposition B.1.

Lemma F.3. *Suppose $\mathsf{K} \in L^2(\mathcal{X}^2, P^2)$ is a characteristic kernel satisfying $\mathbb{E}_{X\sim P}[\mathsf{K}^2(X,X)] < \infty$. Then for $X_1, X_2, \ldots, X_m$ i.i.d. from the distribution P,*

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{m}\sum_{j=1}^{m}\mathsf{K}(X_i,X_j) - \mathbb{E}_{Z\sim P}[\mathsf{K}(X_i,Z)]\right)^2 = 0,$$

on a set $\mathcal{B}_{\mathsf{K}} \in \mathscr{B}(\mathcal{X})$ such that $\mathbb{P}(\mathcal{B}_{\mathsf{K}}) = 1$.

Proof. Let ψ be the feature map corresponding to K. Recalling the definition of mean embedding μ_P from (2.3) observe that, for $1 \leqslant i \leqslant m$,

$$\frac{1}{m}\sum_{j=1}^{m}\mathsf{K}(X_i,X_j) - \mathbb{E}_{Z\sim P}[\mathsf{K}(X_i,Z) = \left\langle \psi(X_i), \frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P \right\rangle_{\mathcal{H}} \tag{F.5}$$

By Cauchy-Schwartz inequality,

$$\left\langle \psi(X_i), \frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P \right\rangle \leqslant \|\psi(X_i)\|_{\mathcal{H}} \left\| \frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P \right\|_{\mathcal{H}} \tag{F.6}$$

By (F.5) and (F.6),

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{m}\sum_{j=1}^{m}\mathsf{K}(X_i,X_j) - \mathbb{E}_{Z\sim P}[\mathsf{K}(X_i,Z)]\right) \\
&\leqslant \left(\frac{1}{m}\sum_{i=1}^{m}\|\psi(X_i)\|_{\mathcal{H}}^2\right)\left\|\frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P\right\|_{\mathcal{H}}^2 \\
&= \left(\frac{1}{m}\sum_{i=1}^{m}\mathsf{K}(X_i,X_i)^2\right)\left\|\frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P\right\|_{\mathcal{H}}^2 .
\end{aligned} \tag{F.7}$$

From (2.3) we have $\mu_P(t) = \mathbb{E}_{X\sim P}[\mathsf{K}(t,X)]$ and hence, by (2.2),

$$\|\mu_P\|_{\mathcal{H}}^2 = \langle \mu_P, \mu_P\rangle_{\mathcal{H}} = \mathbb{E}_{X'\sim P}[\mu_P(X')] = \mathbb{E}_{X,X'\sim P}[\mathsf{K}(X,X')]. \tag{F.8}$$

Once again by (2.3) we observe that,

$$\begin{aligned}
\left\|\frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P\right\|_{\mathcal{H}}^2 &= \left\langle \frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P, \frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P \right\rangle_{\mathcal{H}} \\
&= \frac{1}{m^2}\sum_{1\leqslant i,j\leqslant m}\mathsf{K}(X_i,X_j) - \frac{2}{m}\sum_{j=1}^{m}\mathbb{E}_{Z\sim P}[\mathsf{K}(X_j,Z)] + \|\mu_P\|_{\mathcal{H}}^2 \\
&= \frac{1}{m^2}\sum_{1\leqslant i,j\leqslant m}\mathsf{K}(X_i,X_j) - \frac{2}{m}\sum_{j=1}^{m}\mathbb{E}_{Z\sim P}[\mathsf{K}(X_j,Z)] + \mathbb{E}[\mathsf{K}(X_1,X_2)],
\end{aligned}$$

where the last equality follows from (F.8). Notice that $\mathbb{E}|\mathbb{E}_{Z\sim P}[\mathsf{K}(X_1,Z)]| \leqslant \mathbb{E}|\mathsf{K}(X_1,X_2)| < \infty$. Hence, by the strong law of large numbers for U-statistics [57, Theorem 5.4.A] we conclude that,

$$\lim_{m\to\infty}\left\|\frac{1}{m}\sum_{j=1}^{m}\psi(X_j) - \mu_P\right\|_{\mathcal{H}}^2 = 0, \tag{F.9}$$

on a set $\mathcal{B}_{\mathsf{K}}^{(1)} \in \mathscr{B}(\mathcal{X})$ such that $\mathbb{P}(\mathcal{B}_{\mathsf{K}}^{(1)}) = 1$. Also, since $\mathbb{E}[\mathsf{K}(X_1,X_1)^2] < \infty$, by the strong law of large numbers,

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\mathsf{K}(X_i,X_i)^2 \to \mathbb{E}[\mathsf{K}^2(X_1,X_1)]$$

on a set $\mathcal{B}_{\mathsf{K}}^{(2)}$ such that $\mathbb{P}(\mathcal{B}_{\mathsf{K}}^{(2)}) = 1$. Using this and (F.9) in (F.7) the proof is completed, by choosing $\mathcal{B}_{\mathsf{K}} = \mathcal{B}_{\mathsf{K}}^{(1)} \bigcap \mathcal{B}_{\mathsf{K}}^{(2)}$. □

Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, United States

Email address: anirbanc@wharton.upenn.edu

Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, United States

Email address: bhaswar@wharton.upenn.edu