MMD Aggregated Two-Sample Test

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, Arthur Gretton

Journal of Machine Learning Research 24 (2023) 1-81 Submitted 10/21; Revised 3/23; Published 6/23

MMD Aggregated Two-Sample Test

Antonin Schrab A.SCHRAB@UCL.AC.UK
Centre for Artificial Intelligence, University College London & Inria London
Gatsby Computational Neuroscience Unit, University College London
London, WC1V 6LJ, UK

Ilmun Kim ILMUN@YONSEI.AC.KR
Department of Statistics & Data Science, Department of Applied Statistics, Yonsei University
Seoul, 03722, South Korea

Mélisande Albert MELISANDE.ALBERT@INSA-TOULOUSE.FR
Institut de Mathématiques de Toulouse; UMR 5219, Université de Toulouse; CNRS, INSA; France

Béatrice Laurent BEATRICE.LAURENT@INSA-TOULOUSE.FR
Institut de Mathématiques de Toulouse; UMR 5219, Université de Toulouse; CNRS, INSA; France

Benjamin Guedj B.GUEDJ@UCL.AC.UK
Centre for Artificial Intelligence, University College London & Inria London
London, WC1V 6LJ, UK

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit, University College London
London, W1T 4JG, UK

Editor: Ingo Steinwart

Abstract

We propose two novel nonparametric two-sample kernel tests based on the Maximum Mean Discrepancy (MMD). First, for a fixed kernel, we construct an MMD test using either permutations or a wild bootstrap, two popular numerical procedures to determine the test threshold. We prove that this test controls the probability of type I error non-asymptotically. Hence, it can be used reliably even in settings with small sample sizes as it remains well-calibrated, which differs from previous MMD tests which only guarantee correct test level asymptotically. When the difference in densities lies in a Sobolev ball, we prove minimax optimality of our MMD test with a specific kernel depending on the smoothness parameter of the Sobolev ball. In practice, this parameter is unknown and, hence, the optimal MMD test with this particular kernel cannot be used. To overcome this issue, we construct an aggregated test, called MMDAgg, which is adaptive to the smoothness parameter. The test power is maximised over the collection of kernels used, without requiring held-out data for kernel selection (which results in a loss of test power), or arbitrary kernel choices such as the median heuristic. We prove that MMDAgg still controls the level non-asymptotically, and achieves the minimax rate over Sobolev balls, up to an iterated logarithmic term. Our guarantees are not restricted to a specific type of kernel, but hold for any product of one-dimensional translation invariant characteristic kernels. We provide a user-friendly parameter-free implementation of MMDAgg using an adaptive collection of bandwidths. We demonstrate that MMDAgg significantly outperforms alternative state-of-the-art MMD-based two-sample tests on synthetic data satisfying the Sobolev smoothness assumption, and that, on real-world image data, MMDAgg closely matches the power of tests leveraging the use of models such as neural networks.

Keywords: two-sample testing, kernel methods, minimax adaptivity

Contents

1. Introduction

We consider the problem of nonparametric two-sample testing, where we are given two independent sets of i.i.d. samples, and we want to determine whether these two samples come from the same distribution. This fundamental problem has a long history in statistics and machine learning, with numerous real-world applications in various fields, including clinical laboratory science (Miles et al., 2004), genomics (Chen and Qin, 2010), biology (Fisher et al., 2006), geology (Vermeesch, 2013) and finance (Horváth et al., 2013).

To compare samples from two probability distributions, we use a statistical test of the null hypothesis that the two distributions are equal, against the alternative hypothesis that they are different. Many such tests exist, and rely on different assumptions. If we assume that the two probability distributions are Gaussian with the same variance, then we can perform a Student's t-test (Student, 1908) to decide whether or not to reject the null hypothesis. However, the t-test is parametric in nature, and designed only for comparing two Gaussian distributions. By contrast, our interest is in nonparametric tests, which are sensitive to general alternatives, without relying on specific distributional assumptions. An example of such a nonparametric test is the Kolmogorov–Smirnov test (Massey Jr, 1951) which uses as its test statistic the largest distance between empirical distribution functions of the two samples. The limitation of the Kolmogorov–Smirnov test, however, is that it applies only to univariate data, and its multivariate extension is challenging (Bickel, 1969).

The test statistic we consider is an estimate of the Maximum Mean Discrepancy (MMD—Gretton et al., 2007, 2012a) which is a kernel-based metric on the space of probability distributions. The MMD is an integral probability metric (Müller, 1997) and hence is defined as the supremum, taken over a class of smooth functions, of the difference of their expectations under the two probability distributions. This function class is taken to be the unit ball of a characteristic Reproducing Kernel Hilbert Space (Aronszajn, 1950; Fukumizu et al., 2008; Sriperumbudur et al., 2011), so the Maximum Mean Discrepancy depends on the choice of kernel. We work with a wide range of kernels, each parametrised by their bandwidths.

There exist several heuristics to choose the kernel bandwidths. In the Gaussian kernel case, for example, bandwidths are often simply set to the median distance between pairs of points from both samples (Gretton et al., 2012a). This strategy for bandwidth choice

does not provide any guarantee of optimality, however. In fact, existing empirical results demonstrate that the median heuristic performs poorly (i.e. it leads to low test power) when differences between the two distributions occur at a lengthscale that differs sufficiently from the median inter-sample distance (Gretton et al., 2012b, Figure 1). Ramdas et al. (2015) and Reddi et al. (2015) show that the median heuristic scales as the square root of the dimension, and that using a bandwidth of the higher order with respect to the dimension generally leads to higher power. Another approach is to split the data and learn a good kernel choice on data held out for this purpose (e.g. Gretton et al., 2012b; Liu et al., 2020), however, the resultant reduction in data for testing can reduce overall test power at smaller sample sizes.

Our contributions. Having motivated the problem, we summarize our contributions. We first address the case where the "smoothness parameter" s of the task is known: that is, the distributions being tested have densities in $\mathbb{R}^d$ whose difference lies in a Sobolev ball $\mathcal{S}_d^s(R)$ with smoothness parameter s and radius R. For this setting, we construct a single MMD test that is optimal in the minimax sense over $\mathcal{S}_d^s(R)$, for a specific choice of bandwidths which depend on s.

In practice, s is unknown, and our test must be adaptive to it. We therefore construct a test which is adaptive to s in the minimax sense, by aggregating across tests with different bandwidths, and rejecting the null hypothesis if any individual test (with appropriately corrected level) rejects it. We refer to our proposed MMD aggregated test as MMDAgg. By upper bounding the uniform separation rate of testing of MMDAgg, we prove that it is optimal (up to an iterated logarithmic term) over the Sobolev ball $\mathcal{S}_d^s(R)$ for any $s > 0$ and $R > 0$.

For the practical deployment of our test, we require numerical procedures for computing the test thresholds. We may obtain the threshold for a test of level α using either permutations or a wild bootstrap to estimate the $(1-\alpha)$-quantile of the test statistic distribution under the null hypothesis. We prove that our theoretical guarantees still hold under both test threshold estimation procedures. In the process of establishing these results, we demonstrate the equivalence between using a wild bootstrap and using a restricted set of permutations, which is of independent interest. Through the use of either permutations or a wild bootstrap, we can theoretically guarantee that our proposed single MMD test and MMDAgg test both have non-asymptotic level, which differs from the original MMD test of Gretton et al. (2012a). We believe that this non-asymptotic property of our tests contributes to their real-world applications. Indeed, in practice, settings in which the sample sizes are fixed and may not be assumed to be asymptotically large are very common (*e.g.* medical data, seismological data, data for materials science, *etc.*). While existing tests relying on asymptotic may fail to be well-calibrated in those settings, ours are guaranteed to correctly control the test level non-asymptotically.

We stress that the implementation of MMDAgg corresponds exactly to the test for which we prove theoretical guarantees: we do not make any further approximations in our implementation, nor do we require any prior knowledge on the underlying distribution smoothness. All our theoretical results hold for any product of one-dimensional translation invariant characteristic kernels which are absolutely and square integrable. Our test is, to

the best of our knowledge, the first to be *minimax adaptive* (up to an iterated logarithmic term) for various kernels, and not only for the Gaussian kernel.

Since our approach combines multiple MMD single tests across a large collection of bandwidths, it requires no tuning and our implementation is parameter-free. Furthermore, since we consider various bandwidths simultaneously, our test is adaptive: it performs well both in cases requiring the kernel to have a small bandwidth and in those necessitating a large bandwidth. This means that the same MMDAgg test can detect both local and global differences in densities, which is not the case for a single MMD test with fixed bandwidth.

The key contributions of our paper can be summarised as follows.

- Based on either permutations or a wild bootstrap, we propose a single MMD test and theoretically prove that it has non-asymptotic level. By setting the kernel bandwidth adequately, we show that the test is minimax optimal over a Sobolev ball with known smoothness parameter.

- In order to be adaptive to this smoothness parameter (which is unknown in practice), we construct a two-sample aggregated MMD test, called MMDAgg, which does not require data splitting. We prove that MMDAgg controls the type I error non-asymptotically, and that it is optimal in the minimax sense (up to an iterated logarithmic term) for a wide range of kernels when using either permutations or a wild bootstrap to estimate the test threshold.

- In our experiments on synthetic data, on which the Sobolev smoothness assumption holds, we observe that MMDAgg obtains significantly higher power than all other state-of-the-art MMD adaptive tests considered. On real-world image data, MMDAgg almost matches the power obtained by tests leveraging the capacity of models such as neural networks to detect differences in image distributions. The power of MMDAgg is retained even when a large collection of kernels is considered (up to 12000 kernels in the experiments). Experimentally, no cost in power is incurred for aggregating more kernels, the overall power appears to match the highest power achieved with a single kernel while correctly controlling the test level. As such, in practice, the user can consider as many kernels as computationally feasible.

- We provide a user-friendly parameter-free implementation of MMDAgg, both in Jax and in Numpy, available at `https://github.com/antoninschrab/mmdagg-paper`. This repository also contains code for the reproducibility of our experiments.

Related Works. We present an overview of works related to ours, more details are provided in Section 4. Our non-asymptotic aggregated test, which is minimax adaptive over the Sobolev balls $\left\{\mathcal{S}_d^s(R) : s > 0, R > 0\right\}$, originates from the works of Fromont et al. (2012, 2013) and Albert et al. (2022).

Fromont et al. (2012, 2013) consider the two-sample problem with sample sizes following independent Poisson processes. In this framework, they construct an aggregated test using a different kernel-based estimator with a wild bootstrap. In the multivariate setting, with some condition on the kernel in the Fourier domain, they show minimax adaptivity of their test over Sobolev and anisotropic Nikol'skii-Besov balls. Albert et al. (2022) construct an aggregated independence test using Gaussian kernels. The theoretical guarantees for the

single MMD tests using a permutation-based threshold are related to the result of Kim et al. (2022, Proposition 8.4), also for Gaussian kernels. Besides treating the adaptive two-sample case, rather than the independence case considered by Albert et al. (2022), the present work builds on these earlier results in two important ways. First, the optimality of our aggregated test is not restricted to the use of a specific kernel; it holds more generally for many popular choices of kernels. Second, our theoretical guarantees for this adaptive test are proved to hold even under practical choices for the test threshold: namely the permutation and wild bootstrap approaches.

In this paper, we propose quadratic-time aggregated tests for two-sample testing. This present work, along with those of Albert et al. (2022) and Schrab et al. (2022a) on independence and goodness-of-fit testing, respectively, have been the basis of the later work of Schrab et al. (2022b) who construct efficient (linear-time) variants of these three aggregated tests using incomplete U-statistics, and quantify the trade-off between computational efficiency and test power (in terms of uniform separation rate over Sobolev balls).

Outline. The paper is organised as follows. In Section 2, we formalize the two-sample problem, review the theory of statistical hypothesis testing, and recall the definition of the Maximum Mean Discrepancy. In Section 3, we construct the MMD single and aggregated (MMDAgg) two-sample tests, provide pseudocode for the latter, and derive theoretical guarantees for both. Having introduced the required terminology, we then discuss how our results relate to other works in Section 4. We run various experiments in Section 5 to evaluate how well MMDAgg performs compared to alternative state-of-the-art MMD adaptive tests. The paper closes with discussions and perspectives in Section 6. Proofs, additional discussions, and further experimental results are provided in the Appendices.

2. Background

First, we formalise the two-sample problem in mathematical terms.

Two-sample problem. Given independent samples $\mathbb{X}_m := (X_i)_{1\leq i\leq m}$ and $\mathbb{Y}_n := (Y_j)_{1\leq j\leq n}$, consisting of i.i.d. random variables with respective probability density functions p and q on $\mathbb{R}^d$ with respect to the Lebesgue measure, can we decide whether $p \neq q$ holds?

To tackle this problem, we work in the non-asymptotic framework and construct two nonparametric hypothesis tests in Section 3: a single MMD test for fixed kernel/bandwidth, and MMDAgg which aggregates multiple kernels/bandwidths. In Section 2.1, we first introduce the required notions about hypothesis testing. We then recall the definition of the Maximum Mean Discrepancy and present two estimators for it in Section 2.2.

2.1 Hypothesis testing

We use the convention that $\mathbb{P}_{p\times q}$ denotes the probability with respect to $X_1, \ldots, X_m \overset{\text{iid}}{\sim} p$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} q$ all independent of each other. If given more random variables, say $Z_1, \ldots, Z_t \overset{\text{iid}}{\sim} r$ for some probability density or mass function r, we use the notation $\mathbb{P}_{p\times q\times r}$. We follow similar conventions for expectations and variances.

We address this two-sample problem by testing the null hypothesis $\mathcal{H}_0 \colon p = q$ against the alternative hypothesis $\mathcal{H}_a \colon p \neq q$. Given a *test* Δ which is a function of $\mathbb{X}_m$ and $\mathbb{Y}_n$,

the null hypothesis is rejected if and only if $\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1$. The test is usually designed to control the probability of *type I error*

$$\sup_p \mathbb{P}_{p\times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \leq \alpha$$

for a given $\alpha \in (0,1)$, where the supremum is taken over all probability densities on $\mathbb{R}^d$. We then say that the test has *level* α. For all the definitions, if the test Δ depends on other random variables, we take the probability with respect to those too. For a given fixed level α, the aim is then to construct a test with the smallest possible probability of *type II error*

$$\mathbb{P}_{p\times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0)$$

for specific choices of alternatives for which $p \neq q$. If this probability is bounded by some $\beta \in (0,1)$, we say that the test has *power* $1-\beta$ against that particular alternative. In the asymptotic framework, for a consistent test and a fixed alternative with $\|p-q\|_2 > 0$, we can find large enough sample sizes m and n so that the test has power close to 1 against this alternative. In the non-asymptotic framework of this paper, the sample sizes m and n are fixed. We can then find an alternative with $\|p-q\|_2$ small enough so that the test has power close to 0 against this alternative. Given a test Δ, a class of functions $\mathcal{C}$ and some $\beta \in (0,1)$, one can ask what the smallest value $\tilde{\rho} > 0$ is such that the test Δ has power at least $1-\beta$ against all alternative hypotheses satisfying $p - q \in \mathcal{C}$ and $\|p-q\|_2 > \tilde{\rho}$. Clearly, this depends on the sample sizes: as m and n increase, the value of $\tilde{\rho}$ decreases. This motivates the definition of *uniform separation rate* (Baraud, 2002)

$$\rho(\Delta, \mathcal{C}, \beta, M) \coloneqq \inf\left\{\tilde{\rho} > 0 : \sup_{(p,q)\in\mathcal{F}^M_{\tilde{\rho}}(\mathcal{C})} \mathbb{P}_{p\times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta\right\}$$

where $\mathcal{F}^M_{\tilde{\rho}}(\mathcal{C}) \coloneqq \{(p,q) : \max(\|p\|_\infty, \|q\|_\infty) \leq M, p-q \in \mathcal{C}, \|p-q\|_2 > \tilde{\rho}\}$. For uniform separation rates, we are mainly interested in the dependence on $m+n$: for example we will show upper bounds of the form $a(m+n)^{-b}$ for positive constants a and b independent of m and n. The greatest lower bound on the uniform separation rates of all tests with non-asymptotic level $\alpha \in (0,1)$ is called the *minimax rate of testing* (Baraud, 2002)

$$\underline{\rho}(\mathcal{C}, \alpha, \beta, M) \coloneqq \inf_{\Delta_\alpha} \rho(\Delta_\alpha, \mathcal{C}, \beta, M),$$

where the infimum is taken over all tests Δ_α of non-asymptotic level α for testing $\mathcal{H}_0 \colon p = q$ against $\mathcal{H}_a \colon p \neq q$, and where we compare uniform separation rates in terms of growth rates as functions of $m+n$. This is a generalisation of the concept of critical radius introduced by Ingster (1993a,b) to the non-asymptotic framework. A test is *optimal in the minimax sense* (Baraud, 2002) if its uniform separation rate is upper-bounded up to a constant by the minimax rate of testing. As the class of functions $\mathcal{C}$, we consider the Sobolev ball

$$\mathcal{S}^s_d(R) \coloneqq \left\{f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\widehat{f}(\xi)|^2 \,\mathrm{d}\xi \leq (2\pi)^d R^2\right\} \tag{1}$$

with smoothness parameter $s > 0$, radius $R > 0$, and where $\widehat{f}$ denotes the Fourier transform of f, that is, $\widehat{f}(\xi) \coloneqq \int_{\mathbb{R}^d} f(x) e^{-ix^\top \xi} \,\mathrm{d}x$ for all $\xi \in \mathbb{R}^d$. Our aim is to construct a test which

achieves the minimax rate of testing over $\mathcal{S}_d^s(R)$ (up to an iterated logarithmic term) and which does not depend on the smoothness parameter s of the Sobolev ball; such a test is called *minimax adaptive*.

As shown by Li and Yuan (2019, Theorems 3 and 5), the minimax rate of testing over the Sobolev ball $\mathcal{S}_d^s(R)$ is lower bounded as

$$\underline{\rho}(\mathcal{S}_d^s(R), \alpha, \beta, M) \geq C_0(M, d, s, R, \alpha, \beta)\,(m+n)^{-2s/(4s+d)} \tag{2}$$

for some constant $C_0 > 0$ depending on $\alpha, \beta \in (0,1)$, $d \in \mathbb{N} \setminus \{0\}$ and $M, s, R \in (0, \infty)$. Their proof is an extension of the results of Ingster (1987, 1993b) and we provide more details in Appendix D. We later construct a test with non-asymptotic level α and show in Corollary 7 that its uniform separation rate over $\mathcal{S}_d^s(R)$ with respect to $m+n$ is at most $(m+n)^{-2s/(4s+d)}$, up to some multiplicative constant. This implies that the minimax rate of testing over the Sobolev ball $\mathcal{S}_d^s(R)$ with respect to $m+n$ is exactly of order $(m+n)^{-2s/(4s+d)}$.

2.2 Maximum Mean Discrepancy

As a measure between two probability distributions, we consider the kernel-based *Maximum Mean Discrepancy* (MMD—Gretton et al., 2007, 2012a). In detail, for a given Reproducing Kernel Hilbert Space $\mathcal{H}_k$ (Aronszajn, 1950) with kernel k, the MMD can be formalized as the integral probability metric (Müller, 1997)

$$\mathrm{MMD}(p, q; \mathcal{H}_k) \coloneqq \sup_{f \in \mathcal{H}_k \,:\, \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}_{X\sim p}[f(X)] - \mathbb{E}_{Y\sim q}[f(Y)]|.$$

Our particular interest is in a characteristic kernel k, which guarantees that we have $\mathrm{MMD}(p, q; \mathcal{H}_k) = 0$ if and only if $p = q$. We refer to the works of Fukumizu et al. (2008) and Sriperumbudur et al. (2011) for details on characteristic kernels. It can easily be shown (Gretton et al., 2012a, Lemma 4) that the MMD is the $\mathcal{H}_k$-norm of the difference between the mean embeddings $\mu_p(u) \coloneqq \mathbb{E}_{X\sim p}\left[k(X,u)\right]$ and $\mu_q(u) \coloneqq \mathbb{E}_{Y\sim q}\left[k(Y,u)\right]$ for $u \in \mathbb{R}^d$. Using this fact, a natural unbiased quadratic-time estimator for $\mathrm{MMD}^2(p, q; \mathcal{H}_k)$ (Gretton et al., 2012a, Lemma 6) is

$$\begin{aligned}\widehat{\mathrm{MMD}}_{\mathsf{a}}^2(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{H}_k) \coloneqq \frac{1}{m(m-1)} \sum_{1\leq i\neq i'\leq m} k(X_i, X_{i'}) + \frac{1}{n(n-1)} \sum_{1\leq j\neq j'\leq n} k(Y_j, Y_{j'}) \\ - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j).\end{aligned} \tag{3}$$

This is the minimum variance MMD estimator (Serfling, 1980, Section 5.1.4). As pointed out by Kim et al. (2022), this quadratic-time estimator can be written as a two-sample U-statistic (both of second order) (Hoeffding, 1992)

$$\widehat{\mathrm{MMD}}_{\mathsf{a}}^2(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{H}_k) = \frac{1}{m(m-1)n(n-1)} \sum_{1\leq i\neq i'\leq m} \sum_{1\leq j\neq j'\leq n} h_k(X_i, X_{i'}, Y_j, Y_{j'}) \tag{4}$$

where

$$h_k(x, x', y, y') \coloneqq k(x, x') + k(y, y') - k(x, y') - k(x', y) \tag{5}$$

for $x, y, x', y' \in \mathbb{R}^d$. Writing the estimator $\widehat{\mathrm{MMD}}^2_{\mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{H}_k)$ as a two-sample U-statistic can be theoretically appealing but we stress the fact that it can be computed in quadratic time using Equation (3). The unnormalised version of the test statistic $\widehat{\mathrm{MMD}}^2_{\mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{H}_k)$ was also considered in the work of Fromont et al. (2012).

For the special case when $m = n$, Gretton et al. (2012a, Lemma 6) also propose to consider a different estimator for the Maximum Mean Discrepancy which is the one-sample second-order U-statistic

$$\widehat{\mathrm{MMD}}^2_{\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n; \mathcal{H}_k) \coloneqq \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_k(X_i, X_j, Y_i, Y_j). \tag{6}$$

Note that, unlike the estimator $\widehat{\mathrm{MMD}}^2_{\mathtt{a}}(\mathbb{X}_n, \mathbb{Y}_n; \mathcal{H}_k)$, the estimator $\widehat{\mathrm{MMD}}^2_{\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n; \mathcal{H}_k)$ does not incorporate the terms $\{k(X_i, Y_i) : i = 1, \ldots, n\}$. This means that the ordering of $\mathbb{X}_n = (X_i)_{1 \leq i \leq n}$ and $\mathbb{Y}_n = (Y_j)_{1 \leq j \leq n}$ changes the estimator $\widehat{\mathrm{MMD}}^2_{\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n; \mathcal{H}_k)$. So, when using this estimator, we have to assume we are given a specific ordering of the samples. While $\widehat{\mathrm{MMD}}^2_{\mathtt{b}}$ has slightly higher variance than $\widehat{\mathrm{MMD}}^2_{\mathtt{a}}$, computing $\widehat{\mathrm{MMD}}^2_{\mathtt{b}}$ is computationally much faster than evaluating $\widehat{\mathrm{MMD}}^2_{\mathtt{a}}$, as discussed in Appendix C.

The MMD depends on the choice of kernel, which we explore for our hypothesis tests.

3. Construction of tests and bounds

This section contains our main contributions. In Section 3.1, we introduce some notation along with technical assumptions for our analysis. We then present in Section 3.2 two data-dependent procedures to construct a single MMD test that makes use of some specific kernel bandwidth. Section 3.3 provides sufficient conditions under which this single MMD test is powerful when the difference in densities is measured in terms of the MMD and of the L^2-norm. Based on these preliminary results, we prove an upper bound on its uniform separation rate over the Sobolev ball $\mathcal{S}^s_d(R)$ in Section 3.4, which shows that for a specific choice of bandwidth, the corresponding single MMD test is optimal in the minimax sense. However, the optimal single MMD test relies on the unknown smoothness parameter s of the Sobolev ball $\mathcal{S}^s_d(R)$, which motivates the introduction of our aggregated MMDAgg test in Section 3.5. Finally, we prove in Section 3.6 that MMDAgg is minimax adaptive over the Sobolev balls $\big\{\mathcal{S}^s_d(R) : s > 0, R > 0\big\}$.

3.1 Assumptions and notation

We assume that the sample sizes m and n are balanced up to a constant factor, meaning that there exists a positive constant $C > 0$ such that

$$m \leq n \qquad \text{and} \qquad n \leq Cm. \tag{7}$$

As can be seen in Equation (20), this assumption allows us to upper bound terms such as m^{-1} and n^{-1} by $(m+n)^{-1}$ up to a constant depending on C. The smaller this constant C

is, the tighter our bounds on uniform separation rates will be. For fixed sample sizes, this condition of being balanced is always satisfied. For increasing sample sizes, the condition requires that both sample sizes increase at the same rate. In particular, under this condition, it is not possible to fix one sample size and let the other tend to infinity.

In general, we write $C_i(p_1, \ldots, p_\ell)$ to express the dependence of a positive constant C_i on some parameters $p_1, \ldots, p_\ell$.

We assume that we have d characteristic kernels $(x, y) \mapsto K_i(x - y)$ on $\mathbb{R} \times \mathbb{R}$ for some functions $K_i \colon \mathbb{R} \to \mathbb{R}$ lying in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and satisfying $\int_{\mathbb{R}} K_i(u)\mathrm{d}u = 1$ for $i = 1, \ldots, d$. Then, for some bandwidth $\lambda = (\lambda_1, \ldots, \lambda_d) \in (0, \infty)^d$, the function[1]

$$k_\lambda(x, y) \coloneqq \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$$

is a characteristic kernel on $\mathbb{R}^d \times \mathbb{R}^d$ satisfying[2]

$$\int_{\mathbb{R}^d} k_\lambda(x, y)\mathrm{d}x = 1 \qquad \text{and} \qquad \int_{\mathbb{R}^d} k_\lambda(x, y)^2\mathrm{d}x = \frac{\kappa_2(d)}{\lambda_1 \cdots \lambda_d} \tag{8}$$

for the constant $\kappa_2(d)$ defined later in Equation (21). Using $K_i(u) = \frac{1}{\sqrt{\pi}} \exp\bigl(-u^2\bigr)$ for $u \in \mathbb{R}$ and $i = 1, \ldots, d$, for example, yields the Gaussian kernel k_λ. Using $K_i(u) = \frac{1}{2}\exp(-|u|)$ for $u \in \mathbb{R}$ and $i = 1, \ldots, d$ yields the Laplace kernel k_λ. For notation purposes, given $\lambda = (\lambda_1, \ldots, \lambda_d) \in (0, \infty)^d$, we also write

$$\varphi_\lambda(u) \coloneqq \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{u_i}{\lambda_i}\right) \tag{9}$$

for $u \in \mathbb{R}^d$, so that $k_\lambda(x, y) = \varphi_\lambda(x - y)$ for all $x, y \in \mathbb{R}^d$. In this paper, we investigate the choice of kernel bandwidth for MMD tests. While our theoretical results on test power only hold for translation-invariant kernels on $\mathbb{R}^d \times \mathbb{R}^d$ (as introduced above), we stress that our single and aggregated MMD tests are well-defined and have well-calibrated non-asymptotic levels (Propositions 1 and 8) on any domain and for any positive definite characteristic kernel.

For clarity, we denote $\mathrm{MMD}(p, q; \mathcal{H}_{k_\lambda})$, $\widehat{\mathrm{MMD}}^2_{\mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{H}_{k_\lambda})$, $\widehat{\mathrm{MMD}}^2_{\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n; \mathcal{H}_{k_\lambda})$ and h_{k_λ} (all defined in Section 2) simply by $\mathrm{MMD}_\lambda(p, q)$, $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n)$, $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n)$ and h_λ, respectively.

When $m \neq n$, we let $\widehat{\mathrm{MMD}}^2_{\lambda}(\mathbb{X}_m, \mathbb{Y}_n)$ denote the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n)$. When $m = n$, we let $\widehat{\mathrm{MMD}}^2_{\lambda}(\mathbb{X}_n, \mathbb{Y}_n)$ denote either $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}(\mathbb{X}_n, \mathbb{Y}_n)$ or $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n)$. This means that, when $m = n$, all our results hold for both estimators.

1. Multiplying the kernel k_λ by a positive constant C_λ does not affect the outputs of the single and aggregated tests as it simply scales both the test statistic and the quantile. The requirements that $\int_{\mathbb{R}} K_i(u)\mathrm{d}u = 1$ for $i = 1, \ldots, d$ and the scaling term $(\lambda_1 \cdots \lambda_d)^{-1}$ in the definition of the kernel k_λ are not required for our theoretical results to hold. We introduce those simply for ease of notation in our statements and proofs.
2. Detailed calculations are presented at the beginning of Appendix E.

3.2 Non-asymptotic single MMD test with a fixed bandwidth

Here, we consider the bandwidth $\lambda \in (0,\infty)^d$ to be fixed *a priori*. The null and alternative hypotheses for the two-sample problem are $\mathcal{H}_0 \colon p = q$ against $\mathcal{H}_a \colon p \neq q$, or equivalently $\mathcal{H}_0 \colon \mathrm{MMD}^2_\lambda(p,q) = 0$ against $\mathcal{H}_a \colon \mathrm{MMD}^2_\lambda(p,q) > 0$, provided that the kernels $K_1, \dots, K_d$ are characteristic. Using the samples $\mathbb{X}_m = (X_i)_{1\leq i\leq m}$ and $\mathbb{Y}_n = (Y_j)_{1\leq j\leq n}$, we calculate the test statistic $\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n)$. Since we want our test to be valid in the non-asymptotic framework, we cannot rely on the asymptotic distribution of $\widehat{\mathrm{MMD}}^2_\lambda$ under the null hypothesis to compute the required threshold which guarantees the desired level $\alpha \in (0,1)$. Instead, we use a Monte Carlo approximation to estimate the conditional $(1-\alpha)$-quantile of the permutation-based and wild bootstrap procedures given the samples $\mathbb{X}_m$ and $\mathbb{Y}_n$ under the null hypothesis. For the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}_m, \mathbb{Y}_n)$ defined in Equation (3) we use permutations, while for the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}(\mathbb{X}_n, \mathbb{Y}_n)$ defined in Equation (6) we use a wild bootstrap.

In Appendix B, we provide some more in-depth discussion about the relation between those two procedures. In particular, for the estimate $\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}(\mathbb{X}_n, \mathbb{Y}_n)$, we show in Proposition 11 that using a wild bootstrap corresponds exactly to using permutations which either fix or swap X_i and Y_i for $i = 1, \dots, n$.

3.2.1 Permutation approach

In this case, we consider the MMD estimator defined in Equation (3) which can be written as

$$\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}_m, \mathbb{Y}_n) = \frac{1}{m(m-1)n(n-1)} \sum_{1\leq i\neq i'\leq m} \sum_{1\leq j\neq j'\leq n} h_\lambda(U_i, U_{i'}, U_{m+j}, U_{m+j'})$$

where $U_i \coloneqq X_i$ and $U_{m+j} \coloneqq Y_j$ for $i = 1, \dots, m$ and $j = 1, \dots, n$. Given a permutation function $\sigma \colon \{1, \dots, m+n\} \to \{1, \dots, m+n\}$, we can compute the MMD estimator on the permuted samples $\mathbb{X}^\sigma_m \coloneqq \big(U_{\sigma(i)}\big)_{1\leq i\leq m}$ and $\mathbb{Y}^\sigma_n \coloneqq \big(U_{\sigma(m+j)}\big)_{1\leq j\leq n}$ to get

$$\begin{aligned}
\widehat{M}^\sigma_\lambda &\coloneqq \widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}^\sigma_m, \mathbb{Y}^\sigma_n) \qquad (10)\\
&= \frac{1}{m(m-1)n(n-1)} \sum_{1\leq i\neq i'\leq m} \sum_{1\leq j\neq j'\leq n} h_\lambda(U_{\sigma(i)}, U_{\sigma(i')}, U_{\sigma(m+j)}, U_{\sigma(m+j')})\\
&= \frac{1}{m(m-1)} \sum_{1\leq i\neq i'\leq m} k_\lambda(U_{\sigma(i)}, U_{\sigma(i')}) + \frac{1}{n(n-1)} \sum_{1\leq j\neq j'\leq n} k_\lambda(U_{\sigma(m+j)}, U_{\sigma(m+j')})\\
&\qquad - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k_\lambda(U_{\sigma(i)}, U_{\sigma(m+j)}).
\end{aligned}$$

In order to estimate, with a Monte Carlo approximation, the conditional quantile of $\widehat{M}^\sigma_\lambda$ given $\mathbb{X}_m$ and $\mathbb{Y}_n$, we uniformly sample B i.i.d. permutations $\sigma^{(1)}, \dots, \sigma^{(B)}$. We denote their probability mass function by r, so that $\sigma^{(b)} \sim r$ for $b = 1, \dots, B$. We introduce the notation $\mathbb{Z}_B \coloneqq \big(\sigma^{(b)}\big)_{1\leq b\leq B}$ and also simply write $\widehat{M}^b_\lambda \coloneqq \widehat{M}^{\sigma^{(b)}}_\lambda$ for $b = 1, \dots, B$. We can then use the values $\big(\widehat{M}^b_\lambda\big)_{1\leq b\leq B}$ to estimate the conditional quantile as explained in Section 3.2.3.

3.2.2 Wild bootstrap approach

In this case, we assume that $m = n$ and we work with the MMD estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$ defined in Equation (6). Recall that for this, we must assume an ordering of our samples which gives rise to a pairing (X_i, Y_i) for $i = 1, \dots, n$. Simply using permutations as presented in Section 3.2.1 would break this pairing and our estimators would consist of a signed sum of different terms because

$$\big\{k_\lambda(U_{\sigma(i)}, U_{\sigma(n+i)}) : i = 1, \dots, n\big\} \neq \{k_\lambda(U_i, U_{n+i}) : i = 1, \dots, n\}$$

for most permutations $\sigma\colon \{1, \dots, 2n\} \to \{1, \dots, 2n\}$. The idea is then to restrict ourselves to the permutations σ which, for $i = 1, \dots, n$, either fix or swap X_i and Y_i, in the sense that $\{U_{\sigma(i)}, U_{\sigma(n+i)}\} = \{U_i, U_{n+i}\}$, so that $k_\lambda(U_{\sigma(i)}, U_{\sigma(n+i)}) = k_\lambda(U_i, U_{n+i})$. We show in Proposition 11 in Appendix B that this corresponds exactly to using a wild bootstrap, which we now define.

Given n i.i.d. Rademacher random variables $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n)$ with values in $\{-1, 1\}^n$, we let

$$\widehat{M}^\epsilon_\lambda \coloneqq \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j h_\lambda(X_i, X_j, Y_i, Y_j). \tag{11}$$

As for the permutation approach, in order to obtain a Monte Carlo estimate of the conditional quantile of $\widehat{M}^\epsilon_\lambda$ given $\mathbb{X}_n$ and $\mathbb{Y}_n$, for $b = 1, \dots, B$, we generate n i.i.d. Rademacher random variables $\epsilon^{(b)} \coloneqq \big(\epsilon^{(b)}_1, \dots, \epsilon^{(b)}_n\big)$ with values in $\{-1, 1\}^n$ and compute $\widehat{M}^b_\lambda \coloneqq \widehat{M}^{\epsilon^{(b)}}_\lambda$. We write $\mathbb{Z}_B \coloneqq \big(\epsilon^{(b)}\big)_{1 \leq b \leq B}$ and denote their probability mass function as r to be consistent with the notation introduced in Section 3.2.1, so that $\epsilon^{(b)} \sim r$ for $b = 1, \dots, B$. We next show in Section 3.2.3 how to estimate the conditional quantile using $\big(\widehat{M}^b_\lambda\big)_{1 \leq b \leq B}$.

3.2.3 Single MMD test: definition and level

Depending on which MMD estimator we use, either $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ from Equation (3) or $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$ from Equation (6), we obtain $\big(\widehat{M}^b_\lambda\big)_{1 \leq b \leq B}$ either as in Section 3.2.1 or as in Section 3.2.2, respectively. Inspired by the work of Romano and Wolf (2005a, Lemma 1) and Albert et al. (2022), in order to obtain the prescribed non-asymptotic test level, we also add the original MMD statistic

$$\widehat{M}^{B+1}_\lambda \coloneqq \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n)$$

which corresponds either to the case where the permutation is the identity or where the n Rademacher random variables are equal to 1. We can then estimate the conditional quantile of the distribution of either $\widehat{M}^\sigma_\lambda$ or $\widehat{M}^\epsilon_\lambda$ given $\mathbb{X}_m$ and $\mathbb{Y}_n$ under the null hypothesis $\mathcal{H}_0\colon p = q$ by using a Monte Carlo approximation. In particular, our estimator of the conditional $(1-\alpha)$-quantile is given by

$$\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B \big| \mathbb{X}_m, \mathbb{Y}_n\big) \coloneqq \inf\left\{u \in \mathbb{R} : 1 - \alpha \leq \frac{1}{B+1} \sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}^b_\lambda \leq u\Big)\right\} = \widehat{M}^{\bullet\lceil (B+1)(1-\alpha) \rceil}_\lambda \tag{12}$$

where $\widehat{M}^{\bullet 1}_\lambda \leq \cdots \leq \widehat{M}^{\bullet B+1}_\lambda$ denote the ordered simulated test statistics $\big(\widehat{M}^b_\lambda\big)_{1 \leq b \leq B+1}$. We then define the single MMD test $\Delta^{\lambda,B}_\alpha$ for some given bandwidth $\lambda \in (0, \infty)^d$ as

$$\Delta^{\lambda,B}_\alpha(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_B) \coloneqq \mathbb{1}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B \big| \mathbb{X}_m, \mathbb{Y}_n\big)\Big).$$

Intuitively, $\big(\widehat{M}_\lambda^b\big)_{1\leq b\leq B+1}$ simulate values of the MMD test statistic under the null hypothesis $\mathcal{H}_0\colon p=q$, the quantile $\widehat{q}_{1-\alpha}^{\lambda,B}$ is defined such that only an α-proportion of the simulated test statistics are greater than $\widehat{q}_{1-\alpha}^{\lambda,B}$. As such, as shown in Proposition 1, under $\mathcal{H}_0$, the probability that the MMD test statistic is greater than the quantile (*i.e.* rejecting the null) is non-asymptotically at most α.

As shown in Appendix E.1, the p-value of the test can be computed as

$$p_{\text{val}}^\lambda \coloneqq \frac{1}{B+1}\left(1+\sum_{b=1}^{B}\mathbb{1}\Big(\widehat{M}_\lambda^b \geq \widehat{M}_\lambda^{B+1}\Big)\right)$$

and satisfies the property that

$$p_{\text{val}}^\lambda \leq \alpha \quad\Longleftrightarrow\quad \widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n) > \widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)\,.$$

Note that computing the quantile $\widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)$ requires sorting the simulated test statistics, while computing the p-value p_{val}^λ does not.

We now prove that this single MMD test has the desired non-asymptotic level α, which we stress differs from the original asymptotic MMD test of Gretton et al. (2012a). We believe that this non-asymptotic property will contribute to the wide use of those MMD-based tests.

Proposition 1 (proof in Appendix E.1) *For fixed bandwidth $\lambda \in (0,\infty)^d$, $\alpha\in(0,1)$ and $B\in\mathbb{N}\setminus\{0\}$, the test $\Delta_\alpha^{\lambda,B}$ has non-asymptotic level α, that is*

$$\mathbb{P}_{p\times p\times r}\Big(\Delta_\alpha^{\lambda,B}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_B)=1\Big)\leq\alpha$$

for all probability density functions p on $\mathbb{R}^d$.

This single MMD test $\Delta_\alpha^{\lambda,B}$ depends on the choice of bandwidth λ. In practice, one would like to choose λ such that the test $\Delta_\alpha^{\lambda,B}$ has high power against most alternatives. In general, a smaller bandwidth gives a narrower kernel which is well suited to detect local differences between probability densities such as small perturbations. On the other hand, a larger bandwidth gives a wider kernel which is better at detecting global differences between probability densities. We verify those intuitions in our experiments presented in Section 5. While insightful, those do not tell us exactly how to choose the bandwidth.

As mentioned in the introduction, in practice, there exist two common approaches to choosing the bandwidth of the single MMD test. The first one, proposed by Gretton et al. (2012a), is to set the bandwidth to be equal to the median inter-sample distance. The second approach involves splitting the data into two parts where the first half is used to choose the bandwidth that maximises the asymptotic power, and the second half is used to run the test. This was initially proposed by Gretton et al. (2012b) for the linear-time MMD estimator, and later generalised by Liu et al. (2020) to the case of the quadratic-time MMD estimator. The former approach has no theoretical guarantees, while the latter can suffer from a loss of power caused by the use of less data to run the test. Those two methods are further analysed in our experiments in Section 5.

In Sections 3.3 and 3.4, we obtain theoretical guarantees for the power of the single MMD test $\Delta_\alpha^{\lambda,B}$ and specify the choice of the bandwidth that leads to minimax optimality.

3.3 Controlling the power of the single MMD test

We start by presenting conditions on the discrepancy measures $\mathrm{MMD}_\lambda(p,q)$ and $\|p-q\|_2$ under which the probability of type II error of the single MMD test

$$\mathbb{P}_{p\times q\times r}\Big(\Delta_\alpha^{\lambda,B}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_B)=0\Big)=\mathbb{P}_{p\times q\times r}\Big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)\le \widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)\Big)$$

is controlled by a small positive constant β. We then express these conditions in terms of the bandwidth λ. We find a sufficient condition on the value of $\mathrm{MMD}_\lambda^2(p,q)$ which guarantees that the single MMD test $\Delta_\alpha^{\lambda,B}$ has power at least $1-\beta$ against the alternative $\mathcal{H}_a\colon p\neq q$.

Lemma 2 (proof in Appendix E.2) *For $\alpha,\beta\in(0,1)$, and $B\in\mathbb{N}\setminus\{0\}$, the condition*

$$\mathbb{P}_{p\times q\times r}\left(\mathrm{MMD}_\lambda^2(p,q)\ge\sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)\Big)}+\widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)\right)\ge 1-\frac{\beta}{2}$$

is sufficient to control the probability of type II error such that

$$\mathbb{P}_{p\times q\times r}\Big(\Delta_\alpha^{\lambda,B}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_B)=0\Big)\le\beta.$$

If the densities p and q differ significantly in the sense that $\mathrm{MMD}_\lambda^2(p,q)$ satisfies the condition of Lemma 2, then the probability of type II error of the single MMD test $\Delta_\alpha^{\lambda,B}$ against that alternative hypothesis is upper-bounded by β. The condition includes two terms: the first term depends on β as well as on the variance of $\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)$, and the second is the conditional quantile estimated using the Monte Carlo method with either permutations or a wild bootstrap. In the next two propositions, we make this condition more concrete by providing upper bounds for the variance and the estimated conditional quantile. In particular, the upper bounds are expressed in terms of the bandwidth λ and the sample sizes m and n, which guides us towards the choice of the bandwidth with an optimal guarantee. We start with the variance term.

Proposition 3 (proof in Appendix E.3) *Assume that $\max\left(\|p\|_\infty,\|q\|_\infty\right)\le M$ for some $M>0$. Given φ_λ as defined in Equation (9) and $\psi:=p-q$, there exists a positive constant $C_1(M,d)$ such that*

$$\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)\Big)\le C_1(M,d)\left(\frac{\|\psi*\varphi_\lambda\|_2^2}{m+n}+\frac{1}{(m+n)^2\,\lambda_1\cdots\lambda_d}\right).$$

We now upper bound the estimated conditional quantile $\widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)$ in terms of λ and $m+n$. Since this is a random variable, we provide a bound which holds with high probability.

Proposition 4 (proof in Appendix E.4) *We assume $\max\left(\|p\|_\infty,\|q\|_\infty\right)\le M$ for some $M>0$, $\alpha\in(0,0.5)$ and $\delta\in(0,1)$. For all $B\in\mathbb{N}$ satisfying $B\ge\frac{3}{\alpha^2}\big(\ln\big(\frac{4}{\delta}\big)+\alpha(1-\alpha)\big)$, we have*

$$\mathbb{P}_{p\times q\times r}\left(\widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)\le C_2(M,d)\frac{1}{\sqrt{\delta}}\frac{\ln\big(\frac{1}{\alpha}\big)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\ge 1-\delta$$

for some positive constant $C_2(M,d)$.

Note that while this bound looks similar to the one proposed by Albert et al. (2022, Proposition 3) for independence testing, it differs in two major aspects. Firstly, while they consider the theoretical (unknown) quantile $q^{\lambda}_{1-\alpha}$, we stress that our bound holds for the random variable $\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)$, which is the conditional quantile estimated using the Monte Carlo method with either permutations or a wild bootstrap. Secondly, our bound holds for any bandwidth $\lambda\in(0,\infty)^d$ without any additional assumptions. In particular, we do not require the restrictive condition that $(m+n)\sqrt{\lambda_1\cdots\lambda_d}>\ln\big(\frac{1}{\alpha}\big)$ which can in some cases imply that the sample sizes need to be very large.

Having obtained upper bounds for $\mathrm{var}_{p\times q}\big(\widehat{\mathrm{MMD}}^2_{\lambda}(\mathbb{X}_m,\mathbb{Y}_n)\big)$ and $\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)$, we now combine these with Lemma 2 to obtain a more concrete condition for type II error control. More specifically, the refined condition depends on λ, $m+n$ and β, and guarantees that the probability of type II error of the single MMD test $\Delta^{\lambda,B}_{\alpha}$, against the alternative (p,q) defined in terms of the L^2-norm, is at most β.

Theorem 5 (proof in Appendix E.5) *We assume* $\max\big(\|p\|_\infty,\|q\|_\infty\big)\leq M$ *for some* $M>0$, $\alpha\in(0,e^{-1})$, $\beta\in(0,1)$ *and* $B\in\mathbb{N}$ *which satisfy* $B\geq\frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big)+\alpha(1-\alpha)\big)$. *We consider* φ_λ *as defined in Equation* (9) *and let* $\psi:=p-q$. *Assume that* $\lambda_1\cdots\lambda_d\leq 1$. *There exists a positive constant* $C_3(M,d)$ *such that if*

$$\|\psi\|_2^2\;\geq\;\|\psi-\psi*\varphi_\lambda\|_2^2\;+\;C_3(M,d)\frac{\ln\big(\frac{1}{\alpha}\big)}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}},$$

then the probability of type II error satisfies

$$\mathbb{P}_{p\times q\times r}\Big(\Delta^{\lambda,B}_{\alpha}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_B)=0\Big)\leq\beta.$$

The main condition of Theorem 5 requires $\|p-q\|_2^2$ to be greater than the sum of two quantities. The first one is the bias term $\|\psi-\psi*\varphi_\lambda\|_2^2$ and the second one comes from the upper bounds in Propositions 3 and 4 on the variance $\mathrm{var}_{p\times q}\big(\widehat{\mathrm{MMD}}^2_{\lambda}(\mathbb{X}_m,\mathbb{Y}_n)\big)$ and on the estimated conditional quantile $\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)$. Now, we want to express the bias term $\|\psi-\psi*\varphi_\lambda\|_2^2$ explicitly in terms of the bandwidths λ. For this, we need some smoothness assumption on the difference of the probability densities.

3.4 Uniform separation rate of the single MMD test over a Sobolev ball

We now assume that $\psi:=p-q$ belongs to the Sobolev ball $\mathcal{S}^s_d(R)$ defined in Equation (1). This assumption allows us to derive an upper bound on the uniform separation rate of the single MMD test in terms of the bandwidth λ and of the sum of sample sizes $m+n$.

Theorem 6 (proof in Appendix E.6) *We assume that* $\alpha\in(0,e^{-1})$, $\beta\in(0,1)$, $s>0$, $R>0$, $M>0$ *and* $B\in\mathbb{N}$ *satisfying* $B\geq\frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big)+\alpha(1-\alpha)\big)$. *Given that* $\lambda_1\cdots\lambda_d\leq 1$, *the uniform separation rate of the test* $\Delta^{\lambda,B}_{\alpha}$ *over the Sobolev ball* $\mathcal{S}^s_d(R)$ *can be upper bounded as follows*

$$\rho\Big(\Delta^{\lambda,B}_{\alpha},\mathcal{S}^s_d(R),\beta,M\Big)^2\leq C_4(M,d,s,R,\beta)\left(\sum_{i=1}^d\lambda_i^{2s}+\frac{\ln\big(\frac{1}{\alpha}\big)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)$$

for some positive constant $C_4(M,d,s,R,\beta)$.

The upper bound on the uniform separation rate $\rho\big(\Delta_\alpha^{\lambda,B}, \mathcal{S}_d^s(R), \beta, M\big)$ given by Theorem 6 consists of two terms depending on the bandwidth $\lambda \in (0,\infty)^d$. As the bandwidth λ varies, there is a trade-off between those two quantities: increasing one implies decreasing the other. We can choose the optimal bandwidth λ (depending on $m+n$, d and s) in the sense that both terms have the same order with respect to the sum of sample sizes $m+n$.

Corollary 7 (proof in Appendix E.7) *We assume that $\alpha \in (0, e^{-1})$, $\beta \in (0,1)$, $s > 0$, $R > 0$, $M > 0$ and $B \in \mathbb{N}$ satisfying $B \geq \frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$. The test $\Delta_\alpha^{\lambda^*,B}$ for the choice of bandwidth $\lambda_i^* = (m+n)^{-2/(4s+d)}$, $i = 1, \ldots, d$, is optimal in the minimax sense over the Sobolev ball $\mathcal{S}_d^s(R)$, that is*

$$\rho\Big(\Delta_\alpha^{\lambda^*,B}, \mathcal{S}_d^s(R), \beta, M\Big) \leq C_5(M, d, s, R, \alpha, \beta)\,(m+n)^{-2s/(4s+d)}$$

for some positive constant $C_5(M, d, s, R, \alpha, \beta)$.

We have constructed the single MMD test $\Delta_\alpha^{\lambda^*,B}$ and proved that it is minimax optimal over the Sobolev ball $\mathcal{S}_d^s(R)$ without any restriction on the sample sizes m and n. However, it is worth pointing out that the optimality of the single MMD test hinges on the assumption that the smoothness parameter s is known in advance, which is not realistic. Given this limitation, our next goal is to construct a test which does not rely on the unknown smoothness parameter s of the Sobolev ball $\mathcal{S}_d^s(R)$ and achieves the same minimax rate, up to an iterated logarithmic term, for all $s > 0$ and $R > 0$. This is the main topic of Section 3.5 below.

3.5 Non-asymptotic MMDAgg test aggregating multiple bandwidths

We propose to construct an aggregated test (MMDAgg) by combining multiple single MMD tests, which allows the test to be adaptive to the unknown the smoothness parameter of the Sobolev balls. We use the powerful multiple testing correction of Romano and Wolf (2005b, Equation 9), for which we derive non-asymptotic level guarantees. Consider a finite collection Λ of bandwidths in $(0,\infty)^d$ with an associated collection of positive weights[3] $(w_\lambda)_{\lambda\in\Lambda}$, which will determine the importance of each single MMD test over the others when aggregating all of them. We require that $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$. For notational convenience, we let Λ^w denote the collection of bandwidths Λ with its associated collection of weights. Intuitively, we want to define our aggregated MMDAgg test as the test which rejects the null hypothesis $\mathcal{H}_0 \colon p = q$ if one of the single MMD tests $\big(\Delta_{u_\alpha w_\lambda}^{\lambda,B_1}\big)_{\lambda\in\Lambda}$ rejects the null hypothesis, where u_α is defined as[4]

$$u_\alpha = \sup\left\{u \in \Big(0, \min_{\lambda\in\Lambda} w_\lambda^{-1}\Big) : \mathbb{P}_{p\times p\times r}\left(\max_{\lambda\in\Lambda}\Big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}_{1-uw_\lambda}^{\lambda,B_1}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\Big) > 0\right) \leq \alpha\right\}$$

3. We stress that this differs from the notation often used in the literature (for example, for the independence aggregated test of Albert et al., 2022) where the weights are defined as e^{-w_λ} rather than as w_λ.

4. Since $\alpha \in (0,1)$ and the function $u \mapsto \mathbb{P}_{p\times p\times r}\big(\max_{\lambda\in\Lambda}\big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}_{1-uw_\lambda}^{\lambda,B_1}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\big) > 0\big)$ is non-decreasing, tends to 0 as u tends to 0, and tends to 1 as u tends to $\min_{\lambda\in\Lambda} w_\lambda^{-1}$, u_α is well-defined.

to ensure that MMDAgg has level α. We stress that the data, as well as the choice of collections of bandwidths and weights, all affect the value of u_α. In practice, the probability and the supremum in the definition of u_α cannot be computed exactly. We can estimate the former using a Monte Carlo approximation and estimate the latter using the bisection method. We now explain this in more detail and provide a formal definition of our aggregated MMDAgg test.

For the case of the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ defined in Equation (3), we independently generate a permutation $\sigma^{(b,\ell)} \sim r$ of $\{1,\dots,m+n\}$ and compute $\widehat{M}^b_{\lambda,\ell} \coloneqq \widehat{M}^{\sigma^{(b,\ell)}}_\lambda$ as defined in Equation (10) for $\ell = 1,2$, $b = 1,\dots,B_\ell$ and $\lambda \in \Lambda$. When working with the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$ defined in Equation (6), we independently generate n i.i.d. Rademacher random variables $\epsilon^{(b,\ell)} = \big(\epsilon^{(b,\ell)}_1,\dots,\epsilon^{(b,\ell)}_n\big) \sim r$ and compute $\widehat{M}^b_{\lambda,\ell} \coloneqq \widehat{M}^{\epsilon^{(b,\ell)}}_\lambda$ as defined in Equation (11) for $\ell = 1,2$, $b = 1,\dots,B_\ell$ and $\lambda \in \Lambda$. For consistency between the two procedures, we let $\mathbb{Z}^\ell_{B_\ell} \coloneqq \big(\mu^{(b,\ell)}\big)_{1\le b\le B_\ell}$ for $\ell = 1,2$, where $\mu^{(b,\ell)}$ denotes either the permutation $\sigma^{(b,\ell)}$ or the Rademacher random variable $\epsilon^{(b,\ell)}$ for $\ell = 1,2$ and $b = 1,\dots,B_\ell$. With a slight abuse of notation, we refer to $\mathbb{Z}^1_{B_1}$ and $\mathbb{Z}^2_{B_2}$ simply as $\mathbb{Z}_{B_1}$ and $\mathbb{Z}_{B_2}$. For both estimators, we also let $\widehat{M}^{B_1+1}_{\lambda,1} \coloneqq \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)$. We denote by $\widehat{M}^{\bullet 1}_{\lambda,1} \le \cdots \le \widehat{M}^{\bullet B_1+1}_{\lambda,1}$ the ordered elements $\big(\widehat{M}^b_{\lambda,1}\big)_{1\le b\le B_1+1}$.

We use $\big(\widehat{M}^{\bullet b}_{\lambda,1}\big)_{1\le b\le B_1+1}$, which are computed using $\mathbb{Z}_{B_1}$, $\mathbb{X}_m$ and $\mathbb{Y}_n$, to estimate the conditional $(1-a)$-quantile

$$\widehat{q}^{\lambda,B_1}_{1-a}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big) \coloneqq \widehat{M}^{\bullet\lceil (B_1+1)(1-a)\rceil}_{\lambda,1}$$

for any $a \in (0,1)$ as in Equation (12). As explained in Section 3.2.3, $\widehat{q}^{\lambda,B_1}_{1-a}$ is defined such that an a-proportion of the test statistics $\big(\widehat{M}^b_{\lambda,1}\big)_{1\le b\le B_1+1}$ simulated under the null are greater than $\widehat{q}^{\lambda,B_1}_{1-a}$. By Proposition 1, this ensures the single test with bandwidth λ has non-asymptotic level a.

We use $\big(\widehat{M}^{\bullet b}_{\lambda,2}\big)_{1\le b\le B_2}$, which are computed using $\mathbb{Z}_{B_2}$, $\mathbb{X}_m$ and $\mathbb{Y}_n$, to estimate with a Monte Carlo approximation the probability

$$\mathbb{P}_{p\times q\times r}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) - \widehat{q}^{\lambda,B_1}_{1-uw_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)\right) > 0\right) \tag{13}$$

which appears in the definition of u_α. We denote the approximated quantity by $u^{B_2}_\alpha$, which is formally defined as

$$\begin{aligned}
&u^{B_2}_\alpha\big(\mathbb{Z}_{B_2}\big|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\big)\\
&\coloneqq \sup\left\{u\in\Big(0,\min_{\lambda\in\Lambda} w_\lambda^{-1}\Big): \frac{1}{B_2}\sum_{b=1}^{B_2}\mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2}\Big(\mu^{(b,2)}\Big|\mathbb{X}_m,\mathbb{Y}_n\Big) - \widehat{q}^{\lambda,B_1}_{1-uw_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)\right)>0\right)\le\alpha\right\}\\
&= \sup\left\{u\in\Big(0,\min_{\lambda\in\Lambda} w_\lambda^{-1}\Big): \frac{1}{B_2}\sum_{b=1}^{B_2}\mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2} - \widehat{M}^{\bullet\lceil (B_1+1)(1-uw_\lambda)\rceil}_{\lambda,1}\right) > 0\right) \le \alpha\right\}.
\end{aligned}$$

Since the function $u \mapsto \frac{1}{B_2}\sum_{b=1}^{B_2}\mathbb{1}\Big(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2} - \widehat{M}^{\bullet\lceil (B_1+1)(1-uw_\lambda)\rceil}_{\lambda,1}\right) > 0\Big) - \alpha$ is increasing, $u^{B_2}_\alpha$ is actually the largest root of this function. As such, it can be computed in practice by

using the bisection method for finding the root. We let[5] $\widehat{u}_\alpha^{B_{2:3}} = \widehat{u}_\alpha^{B_{2:3}}\big(\mathbb{Z}_{B_2}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\big)$ be the lower bound of the interval obtained by performing B_3 steps of the bisection method to approximate the supremum (*i.e.* find the root) in the definition of $u_\alpha^{B_2}$. We then have

$$u_\alpha^{B_2} \in \left[\widehat{u}_\alpha^{B_{2:3}},\ \widehat{u}_\alpha^{B_{2:3}} + 2^{-B_3} \min_{\lambda\in\Lambda} w_\lambda^{-1}\right].$$

We recall that the data, the collection of bandwidths, and the weights, all affect the value of the correction u_α, and hence, also the value of its estimate $\widehat{u}_\alpha^{B_{2:3}}$.

For $\alpha \in (0,1)$, we can then define our aggregated MMDAgg test[5] $\Delta_\alpha^{\Lambda^w, B_{1:3}}$ as rejecting the null hypothesis, that is $\Delta_\alpha^{\Lambda^w, B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2}) = 1$, if one of the tests $\left(\Delta^{\lambda, B_1}_{\widehat{u}_\alpha^{B_{2:3}} w_\lambda}\right)_{\lambda\in\Lambda}$ rejects the null hypothesis, that is

$$\exists\, \lambda \in \Lambda : \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\lambda, B_1}_{1-\widehat{u}_\alpha^{B_{2:3}}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\right) w_\lambda}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m, \mathbb{Y}_n\right),$$

or equivalently

$$\exists\, \lambda \in \Lambda : \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{M}^{\bullet\left\lceil (B_1+1)\left(1-\widehat{u}_\alpha^{B_{2:3}}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\right) w_\lambda\right)\right\rceil}_{\lambda,1}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m, \mathbb{Y}_n\right).$$

The parameters of our MMDAgg test $\Delta_\alpha^{\Lambda^w, B_{1:3}}$ are: its level α, the finite collection Λ^w of bandwidths with its associated weights, and the positive integers B_1, B_2 and B_3. We generate independent permutations or Rademacher random variables to obtain $\mathbb{Z}_{B_1}$ and $\mathbb{Z}_{B_2}$. In practice, we are given realisations of $\mathbb{X}_m = (X_i)_{1\le i\le m}$ and $\mathbb{Y}_n = (Y_j)_{1\le j\le n}$. Hence, we are able to compute $\Delta_\alpha^{\Lambda^w, B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2})$ to decide whether or not we should reject the null hypothesis $\mathcal{H}_0\colon p = q$. This exact version of our aggregated MMDAgg test $\Delta_\alpha^{\Lambda^w, B_{1:3}}$ can be implemented in practice with no further approximation. We provide a detailed pseudocode of MMDAgg in Algorithm 1 and our code is available here. In Appendix C, we further discuss how to efficiently compute the values $\widehat{M}^b_{\lambda,\ell}$ for $\ell = 1, 2$, $b = 1, \dots, B_\ell$ and $\lambda \in \Lambda$ (corresponding to Step 1 of Algorithm 1).

The only conditions we have on our weights $(w_\lambda)_{\lambda\in\Lambda}$ for the collection of bandwidths Λ are that they need to be positive and to satisfy $\sum_{\lambda\in\Lambda} w_\lambda \le 1$. We now explain why this condition on the sum of the weights is not necessarily required. In general, the two aggregated tests with weights $(w_\lambda)_{\lambda\in\Lambda}$ and with scaled weights $(w'_\lambda)_{\lambda\in\Lambda}$ where $w'_\lambda := \frac{w_\lambda}{\sum_{\lambda\in\Lambda} w_\lambda}$ for $\lambda \in \Lambda$ are exactly the same. This is due to the way the correction of the levels of the single MMD tests is performed. In particular, making the dependence of $\widehat{u}_\alpha^{B_{2:3}}$ on either Λ^w or $\Lambda^{w'}$ explicit, we have

$$\widehat{u}_\alpha^{B_{2:3}, \Lambda^{w'}}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\right) = \widehat{u}_\alpha^{B_{2:3}, \Lambda^w}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\right) \sum_{\lambda\in\Lambda} w_\lambda,$$

and so $\widehat{u}_\alpha^{B_{2:3}, \Lambda^{w'}}\, w'_\lambda = \widehat{u}_\alpha^{B_{2:3}, \Lambda^w}\, w_\lambda$, which implies that

$$\Delta_\alpha^{\Lambda^{w'}, B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2}) = \Delta_\alpha^{\Lambda^w, B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2}). \tag{14}$$

5. We use the condensed notation $B_{2:3}$ and $B_{1:3}$ to refer to (B_2, B_3) and (B_1, B_2, B_3), respectively.

Algorithm 1: MMDAgg $\Delta_\alpha^{\Lambda^w, B_{1:3}}$

Inputs:

- samples $\mathbb{X}_m = (x_i)_{1\le i\le m}$ in $\mathbb{R}^d$ and $\mathbb{Y}_n = (y_j)_{1\le j\le n}$ in $\mathbb{R}^d$
- choice between permutations (Equation (3)) or wild bootstrap (Equation (6))
- one-dimensional kernels $K_1, \dots, K_d$ satisfying the properties presented in Section 3.1
- level $\alpha \in (0, e^{-1})$
- finite collection of bandwidths Λ in $(0,\infty)^d$
- collection of positive weights $(w_\lambda)_{\lambda\in\Lambda}$ satisfying $\sum_{\lambda\in\Lambda} w_\lambda \le 1$
- number of simulated test statistics B_1 to estimate the quantiles
- number of simulated test statistics B_2 to estimate the level correction
- number of iterations B_3 for the bisection method

Procedure:

Step 1: compute all simulated test statistics (see Appendix C for a more efficient Step 1)

for $\ell = 1, 2$ **and** $b = 1, \dots, B_\ell$:
  generate $\mu^{(b,\ell)} \sim r$ as in Sections 3.2.1 or 3.2.2 (permutations or Rademacher)
  for $\lambda \in \Lambda$:
    compute $\widehat{M}_{\lambda,\ell}^b := \widehat{M}_\lambda^{\mu^{(b,\ell)}}$ as in Equations (10) or (11)

for $\lambda \in \Lambda$:
  compute $\widehat{M}_{\lambda,1}^{B_1+1} := \widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n)$ as in Equations (3) or (6)
  $\left(\widehat{M}_{\lambda,1}^{\bullet 1}, \dots, \widehat{M}_{\lambda,1}^{\bullet B_1+1}\right) = \texttt{sort\_by\_ascending\_order}\left(\widehat{M}_{\lambda,1}^1, \dots, \widehat{M}_{\lambda,1}^{B_1+1}\right)$

Step 2: compute $\widehat{u}_\alpha$ using the bisection method

$u_{\min} := 0$ **and** $u_{\max} := \min_{\lambda\in\Lambda} w_\lambda^{-1}$

repeat B_3 **times**:
  compute $u := \frac{u_{\min}+u_{\max}}{2}$
  compute $P_u := \frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}_{\lambda,2}^b - \widehat{M}_{\lambda,1}^{\bullet\lceil (B_1+1)(1-uw_\lambda)\rceil}\right) > 0\right)$
  if $P_u \le \alpha$ **then** $u_{\min} := u$ **else** $u_{\max} := u$

$\widehat{u}_\alpha := u_{\min}$

Step 3: output test result

if $\widehat{M}_{\lambda,1}^{B_1+1} > \widehat{M}_{\lambda,1}^{\bullet\lceil (B_1+1)(1-\widehat{u}_\alpha w_\lambda)\rceil}$ **for some** $\lambda \in \Lambda$:
  return 1 (reject $\mathcal{H}_0$)
else:
  return 0 (fail to reject $\mathcal{H}_0$)

Time complexity:[6] $\mathcal{O}\big(|\Lambda|(B_1+B_2)(m+n)^2\big)$

Space complexity: $\mathcal{O}\big((m+n)^2 + (B_1+B_2)(m+n)\big)$

6. The time complexity is actually $\mathcal{O}\big(|\Lambda|(B_1+B_2)(m+n)^2 + |\Lambda|B_1\ln(B_1) + |\Lambda|B_2B_3\big)$ which under the reasonable assumption $m+n > \max\big(\sqrt{\ln(B_1)}, \sqrt{B_3}\big)$ gives $\mathcal{O}\big(|\Lambda|(B_1+B_2)(m+n)^2\big)$.

Consider some $u \in \left(0, \min_{\lambda\in\Lambda} w_\lambda^{-1}\right)$. Note that if a single MMD test $\Delta_{uw_\lambda}^{\lambda,B_1}$ has a large associated weight w_λ, then its adjusted level uw_λ is bigger and so the estimated conditional quantile $\widehat{q}_{1-uw_\lambda}^{\lambda,B_1}$ is smaller, which means that we reject this single test more often. Recall that if a single MMD test rejects the null hypothesis, then the aggregated MMDAgg test necessarily rejects the null as well. It follows that a single test $\Delta_{uw_\lambda}^{\lambda,B_1}$ with large weight w_λ is viewed as more important than the other tests in the aggregated procedure. When running an experiment, putting weights on the bandwidths of the single MMD tests can be seen as incorporating prior knowledge about which bandwidths might be better suited to this specific experiment. The choice of prior, or equivalently of weights, is further explored in Section 5.1.

As presented in Section 3.2, the p-value of one of the single MMD tests can be computed as

$$p_{\text{val}}^\lambda := \frac{1}{B_1+1}\left(1 + \sum_{b=1}^{B_1} \mathbb{1}\left(\widehat{M}_{\lambda,1}^b \geq \widehat{M}_{\lambda,1}^{B_1+1}\right)\right)$$

and, with its adjusted level $\widehat{u}_\alpha^{B_{2:3}} w_\lambda$, it satisfies the property that

$$p_{\text{val}}^\lambda \leq \widehat{u}_\alpha^{B_{2:3}} w_\lambda \quad \Longleftrightarrow \quad \widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-\widehat{u}_\alpha^{B_{2:3}} w_\lambda}^{\lambda,B_1}.$$

Hence, our aggregated MMDAgg test can also be expressed in terms of p-values as

$$\begin{aligned}\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2}) &= \mathbb{1}\left(\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-\widehat{u}_\alpha^{B_{2:3}} w_\lambda}^{\lambda,B_1} \text{ for some } \lambda \in \Lambda\right) \\ &= \mathbb{1}\left(p_{\text{val}}^\lambda \leq \widehat{u}_\alpha^{B_{2:3}} w_\lambda \text{ for some } \lambda \in \Lambda\right).\end{aligned}$$

We now show that MMDAgg indeed has non-asymptotic level α. We emphasize the non-asymptotic nature of our aggregated test, which allows for the use of MMDAgg even in settings with small fixed sample sizes, where other asymptotic tests (such as the original MMD test of Gretton et al., 2012a) fail to control correctly the probability of type I error.

Proposition 8 (proof in Appendix E.8) *Consider $\alpha \in (0,1)$ and $B_1, B_2, B_3 \in \mathbb{N}\setminus\{0\}$. For a collection Λ of bandwidths in $(0,\infty)^d$ and a collection of positive weights $(w_\lambda)_{\lambda\in\Lambda}$ satisfying $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$, the MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ has non-asymptotic level α, that is*

$$\mathbb{P}_{p\times p\times r\times r}\left(\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2}) = 1\right) \leq \alpha$$

for all probability density functions p on $\mathbb{R}^d$.

3.6 Uniform separation rate of MMDAgg over Sobolev balls

In this section, we compute the uniform separation rate of our MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ over the Sobolev ball $\mathcal{S}_d^s(R)$. We then present a collection Λ^w of bandwidths and associated weights for which our aggregated test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ is almost optimal in the minimax sense.

First, as part of the proof of Theorem 9 in Equation (25), we have shown that the following bound holds

$$\mathbb{P}_{p\times q\times r\times r}\left(\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2}) = 0\right) \leq \frac{\beta}{2} + \min_{\lambda\in\Lambda} \mathbb{P}_{p\times q\times r}\left(\Delta_{\alpha w_\lambda/2}^{\lambda,B_1}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m, \mathbb{Y}_n\right) = 0\right).$$

This means that we can control the probability of type II error of our MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ by controlling the smallest probability of type II error of the single MMD tests $\big(\Delta^{\lambda,B_1}_{\alpha w_\lambda/2}\big)_{\lambda\in\Lambda}$ with adjusted levels. Hence, given a collection Λ of bandwidths with its associated weights $(w_\lambda)_{\lambda\in\Lambda}$, if for some $\lambda\in\Lambda$ the single MMD test $\Delta^{\lambda,B_1}_{\alpha w_\lambda/2}$ has probability of type II error upper bounded by $\beta/2\in(0,0.5)$, then the probability of type II error of our aggregated MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ is at most β. Intuitively, this means that even if our collection of single MMD tests consists of only one 'good' test (in the sense that it has high power with adjusted level) and many other 'bad' tests (in the sense that they have low power with adjusted levels), MMDAgg would still have high power. This is because when the 'good' MMD test rejects the null hypothesis, MMDAgg also necessarily rejects it. Another point of view on this is that we do not lose any power by testing a wider range of bandwidths as long as the weight of the 'best' test remains the same.

The uniform separation rate of our MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ over the Sobolev ball $\mathcal{S}_d^s(R)$ is then at most twice the lowest of the uniform separation rates of the single MMD tests $\big(\Delta^{\lambda,B_1}_{\alpha w_\lambda/2}\big)_{\lambda\in\Lambda}$. Combining this result with Theorem 6, we obtain the following upper bound on the uniform separation rate of MMDAgg $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ over the Sobolev ball $\mathcal{S}_d^s(R)$.

Theorem 9 (proof in Appendix E.9) *Consider a collection Λ of bandwidths in $(0,\infty)^d$ such that $\lambda_1\cdots\lambda_d\le 1$ for all $\lambda\in\Lambda$ and a collection of positive weights $(w_\lambda)_{\lambda\in\Lambda}$ such that $\sum_{\lambda\in\Lambda} w_\lambda\le 1$. We assume $\alpha\in(0,e^{-1})$, $\beta\in(0,1)$, $s>0$, $R>0$, $M>0$ and $B_1,B_2,B_3\in\mathbb{N}$ satisfying $B_1\ge\big(\max_{\lambda\in\Lambda} w_\lambda^{-2}\big)\frac{12}{\alpha^2}\big(\log\big(\frac{8}{\beta}\big)+\alpha(1-\alpha)\big)$, $B_2\ge\frac{8}{\alpha^2}\ln\big(\frac{2}{\beta}\big)$ and $B_3\ge\log_2\big(\frac{4}{\alpha}\min_{\lambda\in\Lambda}w_\lambda^{-1}\big)$. The uniform separation rate of the aggregated MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ over the Sobolev ball $\mathcal{S}_d^s(R)$ can be upper bounded as follows*

$$\rho\big(\Delta_\alpha^{\Lambda^w,B_{1:3}},\mathcal{S}_d^s(R),\beta,M\big)^2\le C_6(M,d,s,R,\beta)\min_{\lambda\in\Lambda}\left(\sum_{i=1}^d\lambda_i^{2s}+\frac{\ln\big(\frac{1}{\alpha}\big)+\ln\Big(\frac{1}{w_\lambda}\Big)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)$$

for some positive constant $C_6(M,d,s,R,\beta)$.

We recall from Corollary 7 that the optimal choice of bandwidth $\lambda_i^*=(m+n)^{-2/(4s+d)}$, $i=1,\dots,d$, for the single MMD test $\Delta_\alpha^{\lambda^*,B}$ leads to a uniform separation rate over the Sobolev ball $\mathcal{S}_d^s(R)$ of order $(m+n)^{-2s/(4s+d)}$ which is optimal in the minimax sense. However, this choice depends on the unknown smoothness parameter s and so the test cannot be run in practice with this bandwidth. We now propose a specific choice of collection Λ^w of bandwidths and associated weights, which does not depend on s, and derive the uniform separation rate over the Sobolev ball $\mathcal{S}_d^s(R)$ of our aggregated MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ using that collection. Intuitively, the main idea is to construct a collection of bandwidths which includes a bandwidth (denoted λ^*) with the property that

$$\frac{1}{a}\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-2/(4s+d)}\le\lambda_i^*\le\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-2/(4s+d)}$$

for some $a>1$ and for $i=1,\dots,d$. The extra iterated logarithmic term comes from the additional weight term $\ln\big(\frac{1}{w_\lambda}\big)$ in Theorem 9.

Corollary 10 (proof in Appendix E.10) *We assume* $\alpha \in (0, e^{-1})$, $\beta \in (0,1)$, $s > 0$, $R > 0$, $M > 0$, $m + n > 15$ *so that* $\ln(\ln(m+n)) > 1$ *and* $B_1, B_2, B_3 \in \mathbb{N}$ *satisfying* $B_1 \geq \frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$, $B_2 \geq \frac{8}{\alpha^2}\ln\big(\frac{2}{\beta}\big)$ *and* $B_3 \geq \log_2\big(\frac{2\pi^2}{3\alpha}\big)$. *We consider our aggregated MMDAgg test* $\Delta_\alpha^{\Lambda^w, B_{1:3}}$ *with the collection of bandwidths*

$$\Lambda \coloneqq \left\{ \left(2^{-\ell}, \ldots, 2^{-\ell}\right) \in (0,\infty)^d : \ell \in \left\{1, \ldots, \left\lceil \frac{2}{d} \log_2\Big(\frac{m+n}{\ln(\ln(m+n))}\Big) \right\rceil \right\} \right\}$$

and the collection of positive weights $w_\lambda \coloneqq \frac{6}{\pi^2 \ell^2}$ *so that* $\sum_{\lambda \in \Lambda} w_\lambda \leq 1$ *for any sample sizes* m *and* n. *The uniform separation rate of the MMDAgg test* $\Delta_\alpha^{\Lambda^w, B_{1:3}}$ *over the Sobolev ball* $\mathcal{S}_d^s(R)$ *then satisfies*

$$\rho\big(\Delta_\alpha^{\Lambda^w, B_{1:3}}, \mathcal{S}_d^s(R), \beta, M\big) \leq C_7(M, d, s, R, \alpha, \beta) \left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-2s/(4s+d)}$$

for some positive constant $C_7(M, d, s, R, \alpha, \beta)$. *This means that the MMDAgg test* $\Delta_\alpha^{\Lambda^w, B_{1:3}}$, *which does not depend on* s *and* R, *is optimal in the minimax sense up to an iterated logarithmic term over the Sobolev balls* $\big\{\mathcal{S}_d^s(R) : s > 0, R > 0\big\}$; *the MMDAgg test* $\Delta_\alpha^{\Lambda^w, B_{1:3}}$ *is minimax adaptive.*

Note that the choice of using negative powers of 2 for the bandwidths in Corollary 10 is arbitrary. The result holds more generally using negative powers of a for any real number $a > 1$.

With the specific choice of bandwidths and weights of Corollary 10, we have proved that the uniform separation rate of the proposed aggregated MMDAgg test is upper bounded by $((m+n)/\ln(\ln(m+n)))^{-2s/(4s+d)}$. Comparing this with the minimax rate $(m+n)^{-2s/(4s+d)}$, we see that MMDAgg attains rate optimality over the Sobolev ball $\mathcal{S}_d^s(R)$, up to an iterated logarithmic factor, and more importantly, the aggregated test does not depend on the prior knowledge of the smoothness parameter s. Our MMDAgg test is minimax adaptive over the Sobolev balls $\big\{\mathcal{S}_d^s(R) : s > 0, R > 0\big\}$.

4. Related work

In this section, we compare our results to a number of different adaptive kernel hypothesis testing approaches.

Fromont et al. (2012, 2013) construct a two-sample aggregated test in a framework in which the sample sizes follow independent Poisson processes. They use a different kernel-based estimator which corresponds to an unscaled version of the classical quadratic-time MMD estimator of Gretton et al. (2012a, Lemma 6). In their Poisson setting, they derive uniform separation rates for their aggregated test which is minimax adaptive over Sobolev balls, and over anisotropic Nikol'skii-Besov balls, up to an iterated logarithmic term. The quantiles they consider are estimated with a wild bootstrap. They also have an additional assumption on the kernel (condition in the Fourier domain; Fromont et al., 2013, Equation 3.7), which we do not require.

Albert et al. (2022) consider the problem of testing whether two random vectors are dependent and use the kernel-based Hilbert-Schmidt Independence Criterion (HSIC—Gretton

et al., 2005) as a dependence measure. Similarly to our work, they propose a non-asymptotic minimax adaptive test which aggregates single (HSIC) tests, and provide theoretical guarantees: upper bounds for the uniform separation rate of testing over Sobolev and Nikol'skii balls. In their independence testing setting, the information about the problem is encoded in the joint distribution over pairs of variables, with the goal of determining whether this is equal to the product of the marginals. This differs from the two-sample problem we consider, where we have samples from two separate distributions.

Albert et al. (2022) define their single HSIC test using the theoretical quantile of the statistic under the null hypothesis, which is an unknown quantity in practice. To implement the test, they propose a deterministic upper bound on the theoretical quantile (Albert et al., 2022, Proposition 3). This upper bound holds in the two-sample case under the restrictive assumption $(m+n)\sqrt{\lambda_1 \cdots \lambda_d} > \ln\left(\frac{1}{\alpha}\right)$ (this condition is adapted to the two-sample setting from their condition $n\sqrt{\lambda_1 \ldots \lambda_p \mu_1 \ldots \mu_q} > \ln\left(\frac{1}{\alpha}\right)$ for independence testing). If the bandwidth is small (as it can be in the case of the optimal bandwidth λ^* in the proof of Corollary 10), then this condition implies that the results would hold only for very large sample sizes.

By contrast with the above bound, we use a wild bootstrap or permutations to approximate the theoretical quantiles. While the theoretical quantiles are real numbers given data, our estimated quantiles are random variables given data. This means that instead of having a deterministic upper bound on the theoretical quantiles (Albert et al., 2022, Proposition 3), we have an upper bound on our estimated conditional quantiles which holds with high probability as in Proposition 4. Our use of an estimated threshold in place of a deterministic upper bound has an important practical consequence: it allows us to drop the assumption $(m+n)\sqrt{\lambda_1 \cdots \lambda_d} > \ln\left(\frac{1}{\alpha}\right)$ entirely.

Another difference is how the level correction of the single MMD tests is performed. The aggregated test of Albert et al. (2022) involves a theoretical value u_α which cannot be computed in practice, we incorporate directly in our test a Monte Carlo approximation, using either a wild bootstrap or permutations, to estimate the probability under the null hypothesis, and use the bisection method to approximate the supremum. We stress that our theoretical guarantees of minimax optimality (up to an iterated logarithmic term) hold for our aggregated MMDAgg test which can be implemented without any further approximations. Finally, while the results of Albert et al. (2022) hold only for the Gaussian kernel, ours are more general and hold for any product of one-dimensional characteristic translation invariant kernels which are absolutely and square integrable.

Kim et al. (2022, Section 7) propose an adaptive two-sample test for testing equality between two Hölder densities supported on the real d-dimensional unit ball. Instead of testing various bandwidths or kernels, they discretise the support in bins of equal sizes and aggregate tests with varying bin sizes. Each single test is a multinomial test based on the discretised data. Their strategy and the function class they use are both different from the one we consider, but they derive a similar upper bound on the uniform separation rate of testing over Hölder densities. Kim et al. (2022, Proposition 8.4) also mention the setting considered by Albert et al. (2022) and prove an equivalent version of our Theorem 5 for single tests, using permutations for the Gaussian kernel. We consider both the permutation-based and wild bootstrap procedures, and our results hold more generally for a wide range of kernels. With those aforementioned differences, Kim et al. (2022, Example 8.5) anticipate

that one can use a similar reasoning to Albert et al. (2022) to obtain minimax optimality of the single MMD tests. We provide the full statement and proof of this result in our more general setting.

Li and Yuan (2019) present goodness-of-fit, two-sample and independence aggregated asymptotic tests and also establish the minimax rates over Sobolev balls for these three settings. Their tests use the Gaussian kernel and heavily rely on the asymptotic distribution of the test statistic, while our test is non-asymptotic and is not limited to a particular choice of kernel. Their tests are adaptive over Sobolev balls (which they define in a slightly different way than in our case) provided that the smoothness parameter satisfies $s \geq d/4$. We do not have such a restriction. We also note that they assume that the two densities belong to a Sobolev ball, rather than assuming only that the difference of the densities lies in a Sobolev ball. We also point out the work of Tolstikhin et al. (2016) who derive lower bounds for MMD estimation based on finite samples for any radial universal kernel (Sriperumbudur et al., 2011). They establish the minimax rate optimality of the MMD estimators (V-statistic and U-statistic; Lee, 1990).

Gretton et al. (2012b) address kernel adaptation for the linear-time MMD, where the test statistic is computed as a running average (this results in a statistic with greater variance, but allows the processing of larger sample sizes). They propose to choose the kernel by splitting the data, and using one part to select the bandwidth which maximises the estimated ratio of the Maximum Mean Discrepancy to the standard deviation. They show that maximizing this criterion for the linear-time setting corresponds to maximizing the asymptotic power of the test. The test is then performed on the remaining part of the data. Sutherland et al. (2017) and Liu et al. (2020) address kernel adaptation for the quadratic-time MMD using the same sample-splitting strategy, and show that the ratio of the MMD to the standard deviation under the alternative can again be used as a good proxy for test power. Liu et al. (2020) in particular propose a regularized estimator for the variance under the alternative hypothesis, which admits a convenient closed-form expression. Generally, kernel choice by sample splitting gives better results than the median heuristic, as the former is explicitly tuned to optimize the asymptotic power (or a proxy for it). The price to pay for this increase in performance, however, is that we cannot use all the data for the test. In cases where we have access to almost unlimited data this clearly would not be a problem, but in cases where we have a restricted number of samples and work in the non-asymptotic setting, the loss of data to kernel selection might actually result in a net reduction in power, even after kernel adaptation. For better data efficiency, Kübler et al. (2022a) propose an MMD test which uses held-out data not only for kernel selection, but also for choosing weights and test locations for the MMD witness function. In later work, leveraging recent advances in supervised learning and also relying on sample splitting, Kübler et al. (2022b) construct a test which learns the witness function directly by training, for a given amount of time (*i.e.* one minute), an AutoGluon model (Erickson et al., 2020) which can be, for example, a neural network.

Kübler et al. (2020) propose another approach to an MMD adaptive two-sample test which does not require data splitting. Using all the data, they select the linear combination of test statistics with different bandwidths (or even different kernels) which is optimal in the sense that it maximises a power proxy, they then run their test using again all the data. Using the post-selection inference framework (Fithian et al., 2014; Lee et al., 2016), they

are able to correctly calibrate their test to account for the introduced dependencies. This framework requires asymptotic normality of the test statistic under the null hypothesis, however, and hence they are by design restricted to using the linear-time MMD estimate. We observe in our experiments that using this estimate results in a significant loss in power when compared to tests which use the quadratic-time statistic. Yamada et al. (2019) also use post-selection inference to obtain a feature selection method based on the MMD, where the chosen features best distinguish the samples.

In a different setting, Wynne and Duncan (2022) study the efficiency of MMD-based tests when dimension increases, and propose an MMD-based two-sample test for Functional Data Analysis, a framework in which the samples consist of functions rather than of data points. In this setting, Wynne and Nagy (2021) study the connections between kernel mean embeddings and statistical depth (i.e. how representative a point is from a given measure).

5. Experiments

For our aggregated MMDAgg test, we first introduce in Section 5.1 four weighting strategies and a family of collections of bandwidths motivated by Corollary 10. Those collections depend on some parameters which would usually need to be chosen by the user, by contrast, we introduce in Section 5.2 a parameter-free adaptive collection for MMDAgg, which we recommend using in practice. We then present in Section 5.3 some other state-of-the-art MMD-based two-sample tests we will compare ours to. In Section 5.4, we provide details about our experimental procedure. We show that our aggregated MMDAgg test obtains high power on both synthetic and real-world datasets in Sections 5.5 and 5.6, respectively. In Section 5.7, we observe that MMDAgg retains power even in the continuous limit of the collection of bandwidths. We show in Section 5.8 that, on image shift experiments, MMDAgg matches the power of tests using neural networks, even for large sample sizes. Finally, in Section 5.9, we briefly report the results from the additional experiments presented in Appendix A.

5.1 Weighting strategies and fixed bandwidth collections for MMDAgg

The positive weights $(w_\lambda)_{\lambda\in\Lambda}$ for the collection of bandwidths Λ are required to satisfy $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$. As noted in Equation (14), rescaling all the weights to ensure that $\sum_{\lambda\in\Lambda} w_\lambda = 1$ does not change the output of our aggregated test $\Delta_\alpha^{\Lambda^w, B_{1:3}}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}, \mathbb{Z}_{B_2})$. For this reason, we propose weighting strategies for which $\sum_{\lambda\in\Lambda} w_\lambda = 1$ holds.

For any collection Λ of N bandwidths, one can use *uniform* weights which we define as

$$w_\lambda^{\mathrm{u}} \coloneqq \frac{1}{N} \quad \text{for} \quad \lambda \in \Lambda.$$

Using uniform weights should be prioritised if the user does not have any useful prior information to incorporate in the test. The choice of weights is entirely up to the user; the weights can be designed to reflect any given prior belief about the location of the 'best' bandwidths in the collection. Nonetheless, we also present three standard weighting strategies for incorporating prior knowledge when dealing with a more structured collection of bandwidths.

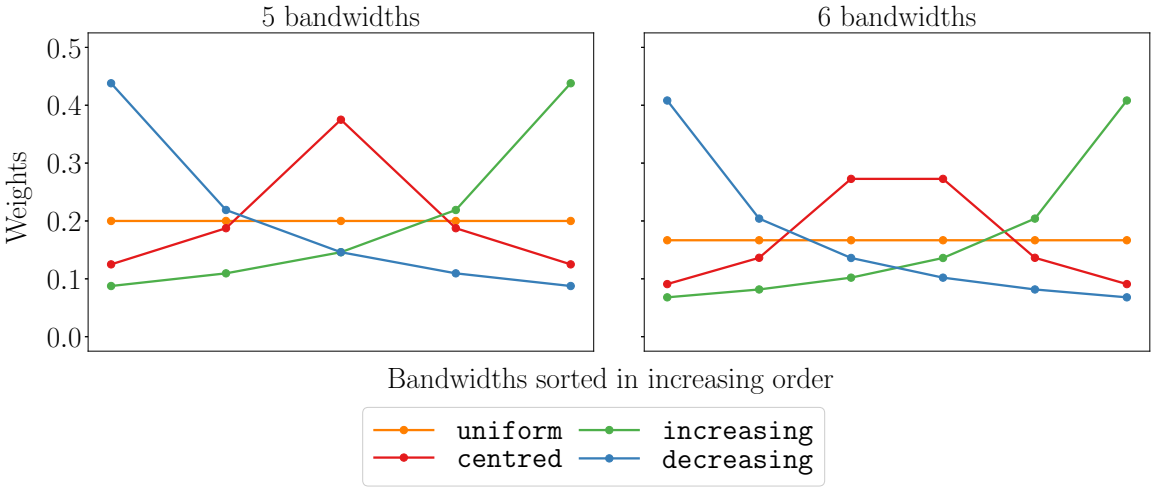

Figure 1: Weighting strategies.

Consider the case where we have some reference bandwidth $\lambda_{ref} \in (0, \infty)^d$ and we are interested in aggregating scaled versions of it, that is, we have an ordered collection of N bandwidths defined as $\lambda^{(i)} := c_i \lambda_{ref}$, $i = 1, \dots, N$, for positive constants $c_1 < \cdots < c_N$. If we have no prior knowledge, then we would simply use the aforementioned uniform weights. Suppose we believe that, if the two distributions differ, then this difference would be better captured by the smaller bandwidths in our collection, in that case we would use *decreasing* weights

$$w^{\mathtt{d}}_{\lambda^{(i)}} := \frac{1}{i} \left(\sum_{\ell=1}^{N} \ell^{-1} \right)^{-1} \quad \text{for} \quad i = 1, \dots, N.$$

On the contrary, if we think that the larger bandwidths in our collection are well suited to capture the difference between the two distributions, if it exists, then we would use *increasing* weights

$$w^{\mathtt{i}}_{\lambda^{(i)}} := \frac{1}{N+1-i} \left(\sum_{\ell=1}^{N} \ell^{-1} \right)^{-1} \quad \text{for} \quad i = 1, \dots, N.$$

Finally, if our prior knowledge is that the bandwidths in the middle of our ordered collection are the most likely to detect the potential difference between the two densities, then we would use *centred* weights which, for N odd, are defined as

$$w^{\mathtt{c}}_{\lambda^{(i)}} := \frac{1}{\left| \frac{N+1}{2} - i \right| + 1} \left(\sum_{\ell=1}^{N} \left(\left| \frac{N+1}{2} - \ell \right| + 1 \right)^{-1} \right)^{-1} \quad \text{for} \quad i = 1, \dots, N,$$

and, for N even, as

$$w^{\mathtt{c}}_{\lambda^{(i)}} := \frac{1}{\left| \frac{N+1}{2} - i \right| + \frac{1}{2}} \left(\sum_{\ell=1}^{N} \left(\left| \frac{N+1}{2} - \ell \right| + \frac{1}{2} \right)^{-1} \right)^{-1} \quad \text{for} \quad i = 1, \dots, N.$$

All those weighting strategies are inspired from the weights of Corollary 10 which are defined as $w_{\lambda^{(i)}} := i^{-2}\big(\sum_{\ell=1}^{\infty} \ell^{-2}\big)^{-1}$ for $i \in \mathbb{N}\backslash\{0\}$, where the square exponent is required in order to have a convergent series. However, in practice, using square exponents in our weights would assign extremely small weights to some of the bandwidths in our collection. This would be almost equivalent to disregarding those bandwidths in our aggregated MMDAgg test, which is not a desired property since if we are not interested in testing some bandwidths, then we would simply not include them in our collection. For this reason, we have defined our weighting strategies without the square exponent. We provide visualisations of our four weighting strategies for collections of 5 and 6 bandwidths in Figure 1.

In our experiments, we refer to our aggregated MMDAgg test with those four weighting strategies as: `MMDAgg uniform`, `MMDAgg decreasing`, `MMDAgg increasing` and `MMDAgg centred`. For these, we use $B_1 = 500$ simulated test statistics to estimate the quantiles, $B_2 = 500$ simulated test statistics to estimate the probability in Equation (13) for the level correction, and $B_3 = 100$ iterations for the bisection method. Motivated by Corollary 10, those tests are used with collections of bandwidths of the form

$$\Lambda(\ell_-, \ell_+) := \Big\{ 2^{\ell} \lambda_{med} \in (0, \infty)^d : \ell \in \big\{\ell_-, \dots, \ell_+\big\} \Big\} \tag{15}$$

for $\ell_-, \ell_+ \in \mathbb{Z}$ such that $\ell_- < \ell_+$, where the median bandwidth λ_{med} is

$$(\lambda_{med})_i := \text{median}\big\{ \big|w_i - w'_i\big| : w, w' \in \mathbb{X}_m \cup \mathbb{Y}_n, w \neq w' \big\}$$

for $i = 1, \dots, d$. We note that, for each experiment, we have chosen the values ℓ_- and ℓ_+ which highlight the differences between the four weighting strategies. In practice, it is not clear how to choose those values, instead, we recommend using the adaptive parameter-free collection of bandwidths introduced in Section 5.2 with uniform weights.

5.2 Adaptive parameter-free collection of bandwidths for MMDAgg

For radial basis function kernels (*i.e.* kernels $k(x, y)$ which can be written as a function of $\|x - y\|$ for some norm $\|\cdot\|$), we recommend using a collection of bandwidths which, intuitively, discretises the interval between the smallest and the largest of the inter-sample distances[7]

$$D := \big\{ \|x - y\| : x \in \mathbb{X}_m, y \in \mathbb{Y}_n \big\}.$$

More formally, we use

$$\Lambda = \left\{ \left(\frac{2\lambda_{max}}{\lambda_{min}/2} \right)^{(i-1)/(N-1)} \lambda_{min}/2 \ : \ i = 1, \dots, N \right\}$$

which is a discretisation of the interval $[\lambda_{min}/2,\, 2\lambda_{max}]$ using N points. We let λ_{min} be the minimum value in D. If this value is smaller than 0.1, we instead use the 5% smallest value in D for λ_{min}, if this quantity is still smaller than 0.1, we use $\lambda_{min} = 0.1$. For λ_{max}, we use the maximum value of D or 0.3 if this maximum is smaller than 0.3. In practice, we recommend using $N = 10$ points, as can be observed in Figure 6 the power remain the same when increasing N to be larger (*i.e.* $N = 100$ or $N = 1000$). Since in general we might not

7. In practice, D can be computed using at most 500 samples from $\mathbb{X}_m$ and 500 samples from $\mathbb{Y}_n$.

have prior information about the location of well suited bandwidths, we recommend using the proposed collection of bandwidths with uniform weights as defined in Section 5.1. We use $B_1 = 2000$ and $B_2 = 2000$ simulated test statistics to estimate the quantiles and the probability in Equation (13) for the level correction, respectively, and use $B_3 = 50$ steps of bisection method. We refer to this test as `MMDAgg`* and emphasize the fact that it is run with exactly the same parameters across all experiments.

5.3 State-of-the-art MMD-based two-sample tests

Gretton et al. (2012a) first suggested using the median heuristic to choose the bandwidth of the MMD test with the Gaussian kernel[8] corresponding to $K_i(u) \coloneqq \frac{1}{\sqrt{\pi}} \exp(-u^2)$ for $u \in \mathbb{R}$, $i = 1, \dots, d$, so that

$$k_\lambda(x, y) \coloneqq \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right) = \frac{1}{\lambda_1 \cdots \lambda_d \pi^{d/2}} \exp\left(-\sum_{i=1}^{d} \left(\frac{x_i - y_i}{\lambda_i}\right)^2\right).$$

They proposed to set the bandwidth to be equal to

$$\lambda_i \coloneqq \text{median}\left\{ \left\| w - w' \right\|_2 : w, w' \in \mathbb{X}_m \cup \mathbb{Y}_n, w \neq w' \right\}$$

for $i = 1, \dots, d$. To generalise this approach to our case where $K_1, \dots, K_d$ need not all be the same, as explained in Section 5.1, we can in a similar way set the bandwidth[9] to

$$\lambda_i \coloneqq \text{median}\left\{ \left| w_i - w_i' \right| : w, w' \in \mathbb{X}_m \cup \mathbb{Y}_n, w \neq w' \right\}$$

for $i = 1, \dots, d$. With this specific definition for the bandwidth, we use the notation $\lambda_{med} \coloneqq (\lambda_1, \dots, \lambda_d)$. We refer to the single MMD test with the bandwidth λ_{med} as `median` in our experiments.

Another common approach for selecting the bandwidth was first introduced by Gretton et al. (2012b) for the single MMD test using the linear-time MMD estimator. The method was then extended to the case of the quadratic-time MMD estimator by Sutherland et al. (2017). It consists in splitting the data in two parts and in using the first part to select the bandwidth which maximises the asymptotic power of test, or equivalently the estimated ratio (Liu et al., 2020, Equations 4 and 5)

$$\frac{\widehat{\text{MMD}}^2_\lambda(\mathbb{X}_n, \mathbb{Y}_n)}{\widehat{\sigma}_\lambda(\mathbb{X}_n, \mathbb{Y}_n)} \tag{16}$$

where

$$\widehat{\sigma}^2_\lambda(\mathbb{X}_n, \mathbb{Y}_n) \coloneqq \frac{4}{n^3} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} h_\lambda(X_i, X_j, Y_i, Y_j) \right)^2 - \frac{4}{n^4} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} h_\lambda(X_i, X_j, Y_i, Y_j) \right)^2 + 10^{-8}$$

8. Gretton et al. (2012a) actually consider the unnormalised Gaussian kernel without the $\left(\lambda_1 \cdots \lambda_d \pi^{d/2}\right)^{-1}$ term, but as pointed out in Footnote 1, this does not affect the output of the test.
9. Note that those two ways of setting the bandwidths are not equivalent for the Gaussian kernel but they are each equally valid.

is a regularised positive estimator of the asymptotic variance of the quadratic-time estimator $\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_n, \mathbb{Y}_n)$ under the alternative hypothesis $\mathcal{H}_a$ for $m = n$. In our experiments, we select the bandwidth of the form $c\lambda_{med}$ for various positive values of c which maximises the estimated ratio. The single MMD test with the selected bandwidth is then performed on the second part of the data. In our experiments, we refer to this test as `split`.

Another interesting test to compare ours to, is the one which uses new data to choose the optimal bandwidth. This corresponds to running the above test which uses data splitting on twice the amount of data. In some sense, this represents the best performance the single MMD test can achieve as the test is run on the whole dataset with an optimal choice of bandwidth. As such, it is interesting to observe the difference in power between MMDAgg and this oracle test which uses extra data. In our experiments, we denote this test as `oracle`.

A radically different approach to constructing an MMD adaptive two-sample test was recently presented by Kübler et al. (2020). They work in the asymptotic regime and require asymptotic normality under the null hypothesis of their MMD estimator, so they are restricted to using the linear-time estimator. Using all of the data, they compute the linear-time MMD estimates for several kernels (or several bandwidths of a kernel), they then select the linear combination of these which maximises a proxy for asymptotic power, and compare its value to their test threshold. They do not split the data but they are able to correctly calibrate their test for the introduced dependencies. For this, they prove and use a generalised version of the post-selection inference framework (Fithian et al., 2014; Lee et al., 2016) which holds for uncountable candidate sets (i.e. all linear combinations). In our experiments, we compare our aggregated MMDAgg test to their one-sided test (OST—Kübler et al., 2020) for which we use their implementation. This test is referred to as `ost` in our experiments.

In later work, Kübler et al. (2022b) propose an AutoML (Automated Machine Learning) test with an implementation which is essentially parameter-free. Their test relies on sample splitting, on cross-validation, and on permuting the data, the witness function of the test is learnt by training an AutoGluon model (Erickson et al., 2020) for some prescribed amount of time (one minute by default). Depending on the time limit and on the compute available, a different model will be chosen automatically. While such a black-box approach can certainly be convenient for everyday users, this convenience comes at the expense of reproducibility (even on the same machine it can take different amounts of time to train identical models). We refer to this test in our experiments as `AutoML`.

5.4 Experimental procedure

To compute the median bandwidth

$$(\lambda_{med})_i \coloneqq \mathrm{median}\left\{\left|w_i - w'_i\right| : w, w' \in \mathbb{X}_m \cup \mathbb{Y}_n, w \neq w'\right\}$$

for $i = 1, \ldots, d$, we use at most 1000 randomly selected data points from $\mathbb{X}_m$ and at most 1000 from $\mathbb{Y}_n$, since the median is robust, this is sufficient to get an accurate estimate of it. Moreover, we use a threshold so that the bandwidth is not smaller than 0.0001. This avoids division by 0 in some settings where one component of the data points is always the

same value, as it can be the case for the problem considered in Section 5.6 which uses the MNIST dataset, where the pixel in one corner of the images is always black for every digit.

We run all our experiments with the Gaussian kernel

$$k_\lambda(x,y) \coloneqq \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right) = \frac{1}{\lambda_1 \cdots \lambda_d \pi^{d/2}} \exp\left(-\sum_{i=1}^{d}\left(\frac{x_i - y_i}{\lambda_i}\right)^2\right)$$

for $K_i(u) \coloneqq \frac{1}{\sqrt{\pi}} \exp(-u^2)$ for $u \in \mathbb{R}$, $i = 1, \ldots, d$, and with the Laplace kernel

$$k_\lambda(x,y) \coloneqq \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right) = \frac{1}{\lambda_1 \cdots \lambda_d 2^d} \exp\left(-\sum_{i=1}^{d}\left|\frac{x_i - y_i}{\lambda_i}\right|\right)$$

for $K_i(u) \coloneqq \frac{1}{2} \exp(-|u|)$ for $u \in \mathbb{R}$, $i = 1, \ldots, d$. As mentioned in Footnote 1, MMDAgg does not depend on the scaling of the kernels. Hence, in our implementation we drop the scaling terms in front of the exponential functions, which is numerically more stable.

We use a wild bootstrap for all our experiments, except for the one in Appendix A.3 where we compare using the permutation-based and wild bootstrap procedures, and for the one in Appendix A.4 where we must use permutations as we consider different sample sizes $m \neq n$. We use level $\alpha = 0.05$ for all our experiments.

We run all our tests on three different types of data: 1-dimensional and 2-dimensional perturbed uniform distributions, and the MNIST dataset. Those are introduced in Sections 5.5 and 5.6, respectively. We use sample sizes $m = n = 500$ for the 1-dimensional perturbed uniform distributions and for the MNIST dataset, and use larger sample sizes $m = n = 2000$ for the case of the 2-dimensional perturbed uniform distributions.

For the `split` and `oracle` tests, we use two equal halves of the data, and `oracle` is run on twice the sample sizes. We choose the bandwidth which maximises the estimated ratio presented in Equation (16) out of the collection $\{c\lambda_{med} : c \in \{0.1, 0.2, \ldots, 0.9, 1\}\}$ when considering perturbed uniform distributions, and when considering the MNIST dataset we select it out of the collection $\{2^c \lambda_{med} : c \in \{10, 11, \ldots, 19, 20\}\}$. Similarly to our aggregated tests of Section 5.1, for the `median`, `split` and `oracle` tests, we use $B = 500$ simulated test statistics to estimate the quantile.

For `MMDAgg`*, we also use either the Gaussian or the Laplace kernel with $N = 10$ bandwidths, we refer to those tests as `MMDAgg`* `Gaussian` and `MMDAgg`* `Laplace`. We also consider aggregating over different types of kernels, in particular, we can use both the Gaussian and Laplace kernels, each with $N = 10$ bandwidths. This gives a collection consisting of $2N = 20$ kernels, over which `MMDAgg`* `Laplace Gaussian` aggregates. Finally, we propose to aggregate 12 kernels (each with $N = 10$ bandwidths): the Gaussian kernel, the inverse multiquadric (IMQ) kernel, and the Matérn kernels with the ℓ^1 and ℓ^2 distances for $\nu = 0.5, 1.5, 2.5, 3.5, 4.5$ (the Laplace kernel is the Matérn kernel with ℓ^1 distance and $\nu = 0.5$). This test aggregates over $12N = 120$ kernels, we refer to it as `MMDAgg`* `All`. Note that in Figure 6, we consider $N = 1000$ bandwidths, which means for example that the `MMDAgg`* `All` aggregates over $12N = 12000$ kernels, and we observe that it retains its high power.

To estimate the power in our experiments, we average the test outputs of 500 repetitions, that is, 500 times, we sample some new data and run the test. We sample new data for each

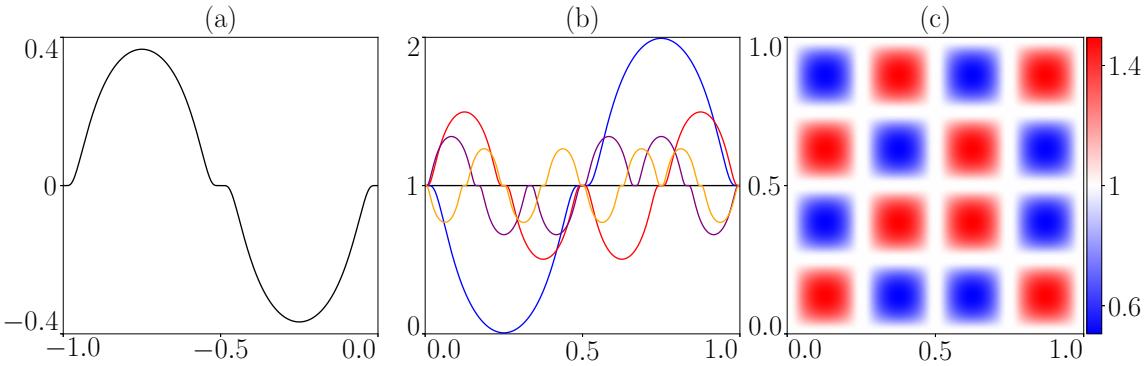

Figure 2: (a) Function G, (b) 1-dimensional uniform distribution with 0, 1, 2, 3 and 4 perturbations, (c) 2-dimensional uniform distribution with 2 perturbations.

test with different parameters, except when we compare using either a wild bootstrap or permutations, in which case we use the same samples. All our experiments are reproducible and our code is available here.

5.5 Power experiments on synthetic data

As explained in Appendix D, a lower bound on the minimax rate of testing over the Sobolev ball $\mathcal{S}_d^s(R)$ can be obtained by considering a d-dimensional uniform distribution and a perturbed version of it with $P \in \mathbb{N} \setminus \{0\}$ perturbations. As presented in Equation (19), the latter has density

$$f_\theta(u) := \mathbb{1}_{[0,1]^d}(u) + c_d P^{-s} \sum_{\nu \in \{1,\dots,P\}^d} \theta_\nu \prod_{i=1}^{d} G\left(Pu_i - \nu_i\right), \quad u \in \mathbb{R}^d \tag{17}$$

where $\theta = (\theta_\nu)_{\nu \in \{1,\dots,P\}^d} \in \{-1,1\}^{P^d}$, that is, θ is a vector of length P^d with entries either -1 or 1, and it is indexed by the P^d d-dimensional elements of $\{1,\dots,P\}^d$, and

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

We have added a scaling factor c_d to emphasize the effect of the perturbations, in our experiments we use $c_1 = 2.7$ and $c_2 = 7.3$. Those values were chosen to ensure that the densities with one perturbation remain positive on $[0,1]^d$. The uniform density with P perturbations for $P = 0, 1, 2, 3, 4$ when $d = 1$ and for $P = 2$ when $d = 2$, as well as the function G, are plotted in Figure 2. As shown by Li and Yuan (2019), for P large enough, the difference between the uniform density and the perturbed uniform density lies in the Sobolev ball $\mathcal{S}_d^s(R)$ for some $R > 0$. In our experiments, we choose the smoothness parameter of the perturbed uniform density defined in Equation (17) to be equal to $s = 1$. For each of the 500 repetitions used to estimate the power of a test, we sample uniformly a new value of the parameter $\theta \in \{-1,1\}^{P^d}$ for the perturbed uniform density.

In Figure 3, we consider testing $n = 500$ samples drawn from the 1-dimensional uniform distribution against $m = 500$ samples drawn from a 1-dimensional uniform distribution

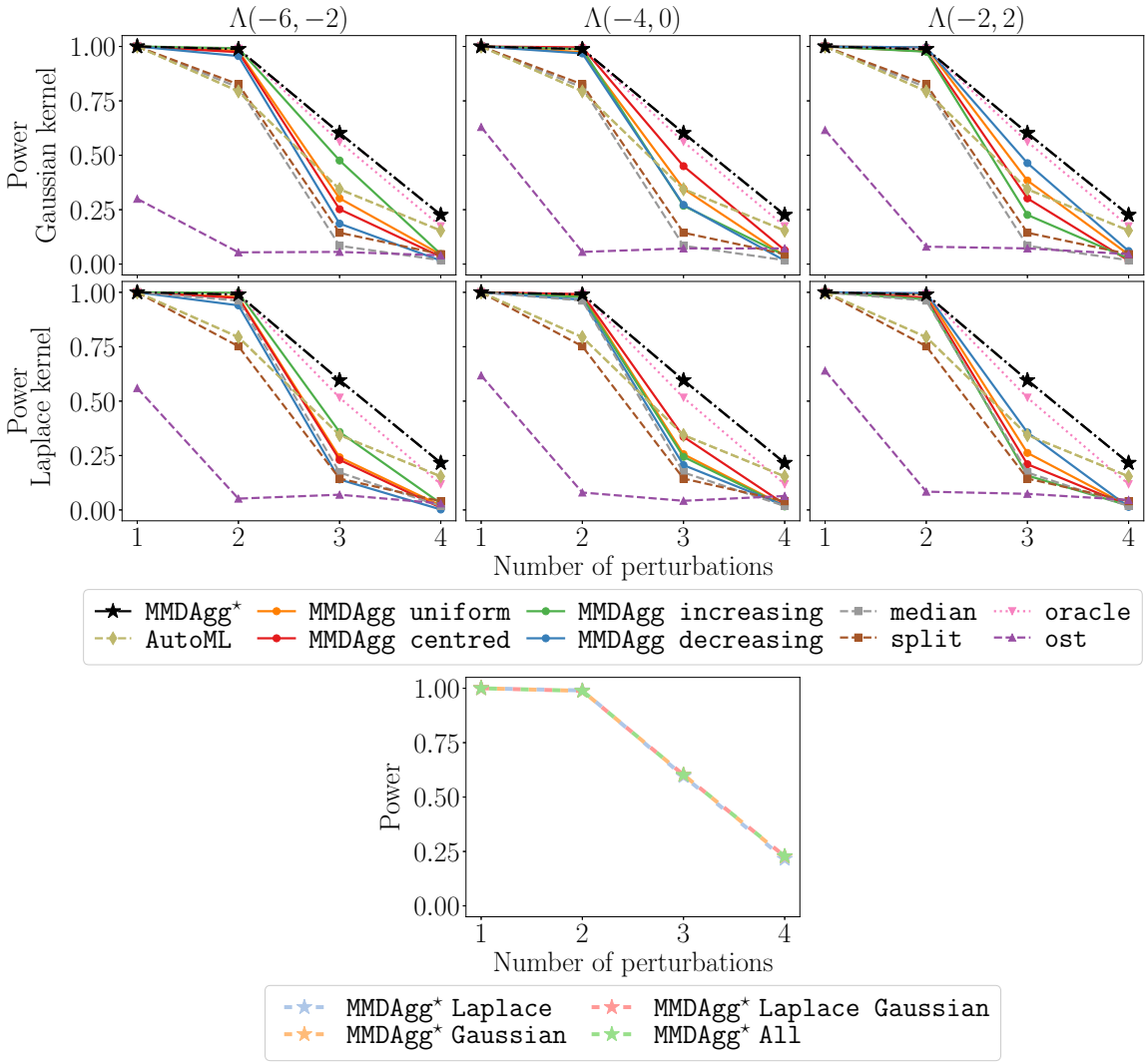

Figure 3: Power experiments with 1-dimensional perturbed uniform distributions using sample sizes $m = n = 500$ with a wild bootstrap.

with $P = 1, 2, 3, 4$ perturbations. We consider the same setting but in two dimensions with sample sizes $m = n = 2000$ with up to three perturbations in Figure 4. For the MMDAgg tests of Section 5.1 and for the `ost` test, we use the collections of bandwidths $\Lambda(-6, -2)$, $\Lambda(-4, 0)$ and $\Lambda(-2, 2)$ as defined in Equation (15). As the number of perturbations increases, it becomes harder to distinguish the two distributions, this translates into a decrease in power for all the tests. Even though we consider more samples for the 2-dimensional case, the performance of all the tests degrades significantly with dimension as detecting the perturbations becomes considerably more challenging.

For all the settings considered in Figures 3 and 4, we observe that `MMDAgg*` always performs the best, with power slightly higher than the one of `oracle` which has access to extra data to select an optimal bandwidth. All other tests achieve significantly lower test

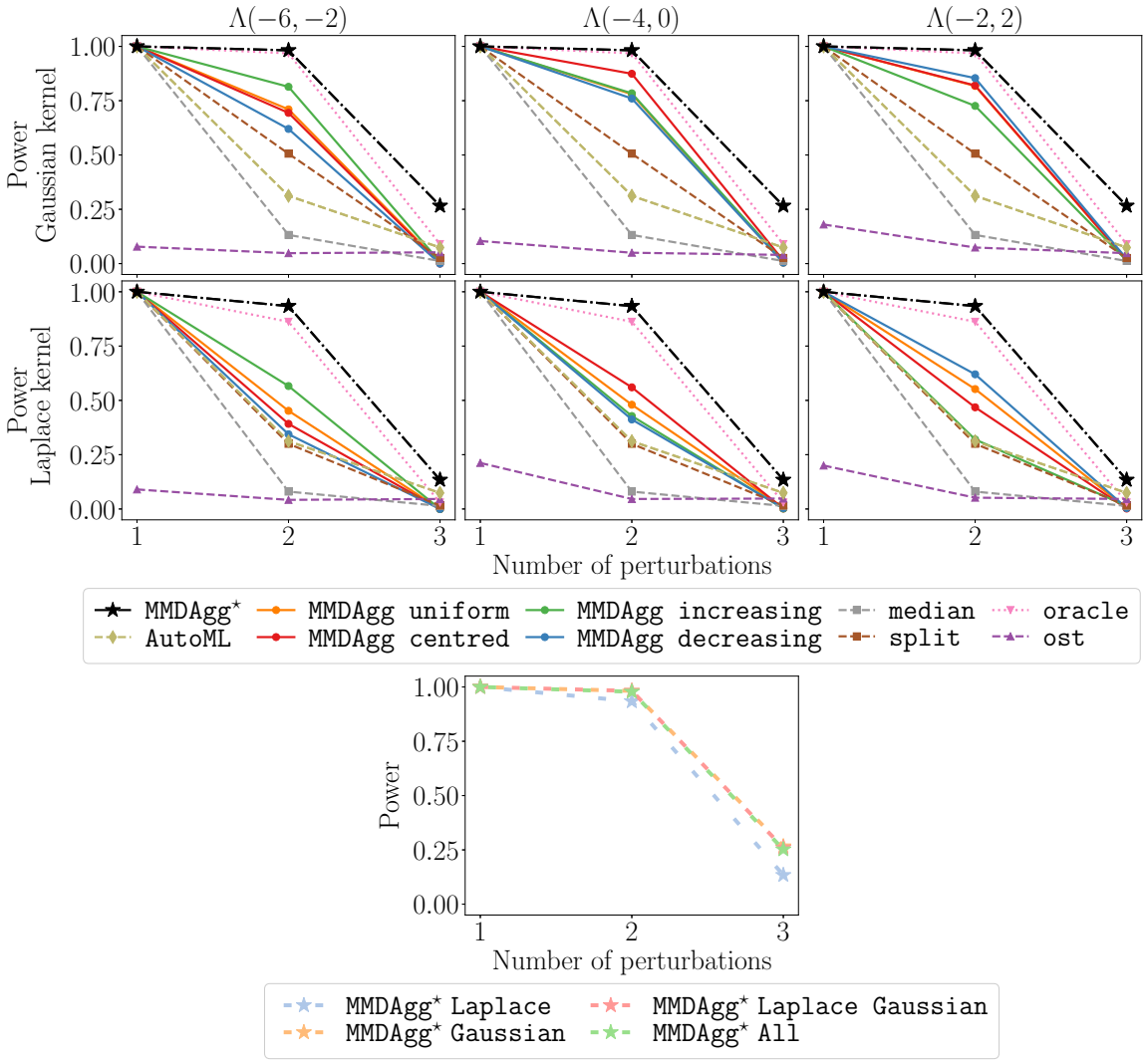

Figure 4: Power experiments with 2-dimensional perturbed uniform distributions using sample sizes $m = n = 2000$ with a wild bootstrap.

power. This validates our theoretical results that our aggregated MMDAgg test is minimax optimal in settings such as the one considered in this experiment where the difference in densities lies in a Sobolev ball of unknown smoothness. In particular, with its adaptive collection of bandwidths, `MMDAgg*` obtains higher power than the four other aggregated tests with specifically chosen collections. In the 1-dimensional case, `AutoML` performs similarly to our four aggregated tests, while in the 2-dimensional case with two perturbations it obtains much lower power. Our four tests with different weighting strategies outperform the three other tests `median`, `split` and `ost` in most settings, and always at least match the power of the best of those three. The `ost` test, which is restricted to using the linear-time MMD estimate, obtains very low power compared to all the other tests using the quadratic-time estimate.

In all the experiments in Figures 3 and 4, using the Gaussian rather than the Laplace kernel results in higher power for our four aggregated tests, the difference is small but notable for the 1-dimensional case while it is large for the 2-dimensional case. For `MMDAgg`*, there is no difference in Figure 3 and a small one in Figure 4, as can be seen in the bottom plots. In one dimension, the `median` test performs significantly better when using the Laplace kernel and outperforms the `split` test, which is not the case in all the other settings.

In the bottom plots of Figures 3 and 4, we observe that by aggregating over both Gaussian and Laplace kernels, `MMDAgg`* `Laplace Gaussian` obtains the highest power achieved by either `MMDAgg`* `Laplace` or `MMDAgg`* `Gaussian`. When adding 10 other kernels to the collection (each with 10 bandwidths), `MMDAgg`* `All` retains the same high power, we do not observe a cost in power for considering more kernels. This is only possible due the way we perform the level correction in Section 3.5.

We now discuss the relation between the four weighting strategies for MMDAgg. Recall from Section 3.5 that a single MMD test with larger associated weight is viewed as more important than one with smaller associated weight in the aggregated procedure. Recall from Section 5.1 that `MMDAgg uniform` puts equal weights on every bandwidths, that `MMDAgg centred` puts the highest weight on the bandwidth in the middle of the collection, and that `MMDAgg increasing` puts the highest weight on the biggest bandwidth while `MMDAgg decreasing` puts it on the smallest bandwidth. This allows us to interpret our results.

First, let's consider the case of the collection of bandwidths $\Lambda(-6, -2)$ for both one and two dimensions. We observe that `MMDAgg increasing` has the highest power and `MMDAgg decreasing` the lowest of the four aggregated tests, this means that putting the highest weight on the biggest bandwidth performs the best while putting it on the smallest bandwidth performs the worst. We can deduce that the most important bandwidth in our collection is the biggest one, which suggests that we should consider a collection consisting of larger bandwidths, say $\Lambda(-4, 0)$. In this case, `MMDAgg centred` now obtains the highest power of our four weighting strategies. We can infer that the optimal bandwidth is close to the bandwidths in the middle of our collection. When considering a collection of even larger bandwidths $\Lambda(-2, 2)$, we see the opposite trends to ones observed using $\Lambda(-6, -2)$; `MMDAgg decreasing` and `MMDAgg increasing` are performing the best and worst of our four tests, respectively. This suggests that a collection consisting of smaller bandwidths than $\Lambda(-2, 2)$ might be more appropriate.

So, comparing our aggregated tests with different weighting strategies gives us some insights on whether the collection we have considered is appropriate, or consists of bandwidths which are either too small or too large. The uniform weighting strategy does not perform the best but it is more robust to changes in the collection of bandwidths than the other strategies. Of course, in practice, if we have access to a limited amount of data, one cannot run a hypothesis test with some parameters, observe the results and then modify those parameters to run the test again. Nonetheless, the interpretation of the results of our different weighting strategies remains an appealing feature of our tests. In practice, we recommend using the parameter-free test `MMDAgg`* `Laplace Gaussian` with its collection of bandwidths chosen adaptively.

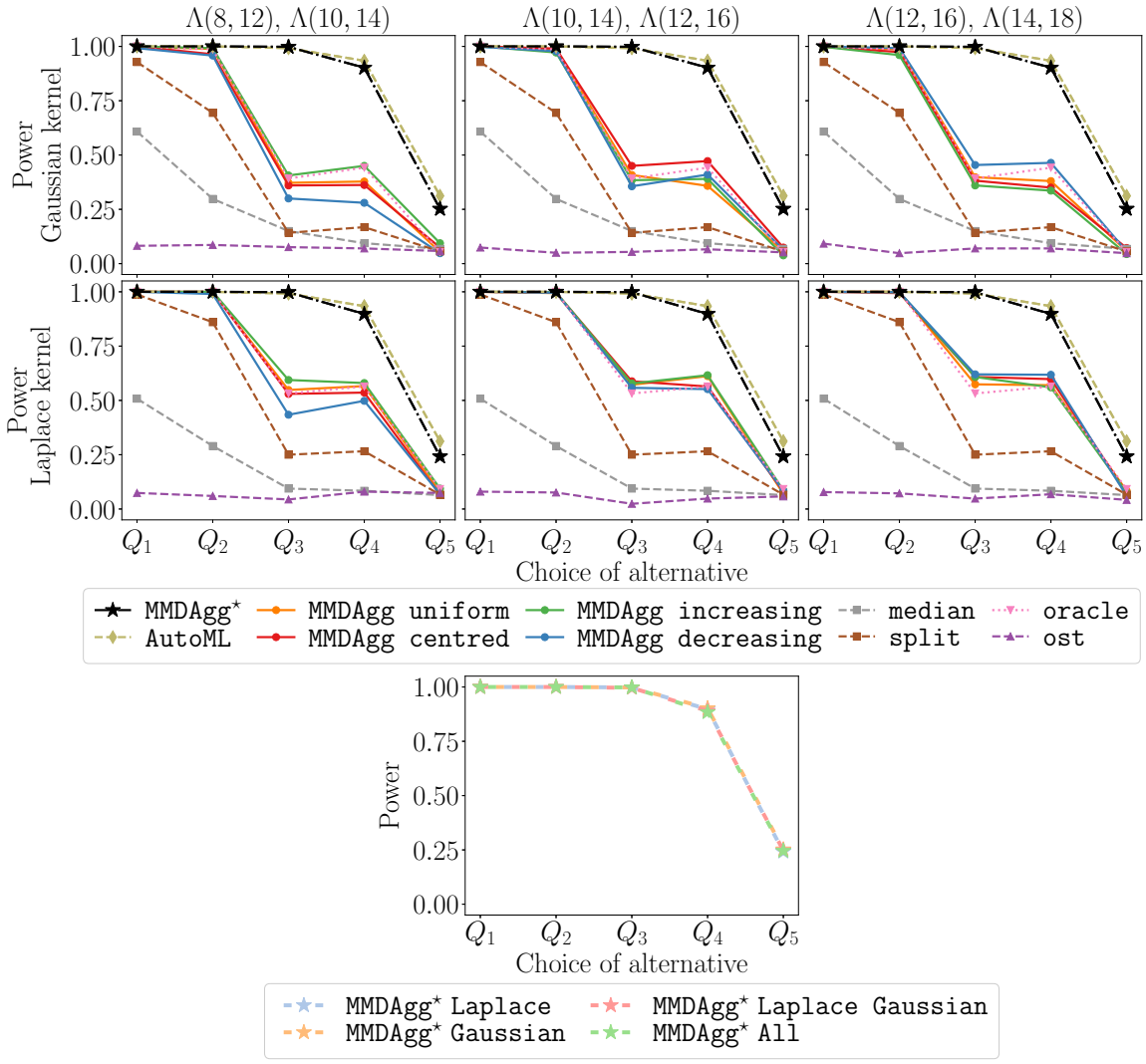

Figure 5: Power experiments with the MNIST dataset using sample sizes $m = n = 500$ with a wild bootstrap.

5.6 Power experiments on the MNIST dataset

Motivated by the experiment considered by Kübler et al. (2020), we consider the MNIST dataset (LeCun et al., 2010) down-sampled to 7×7 images. In Figure 5, we consider 500 samples drawn with replacement from the set $\mathcal{P}$ consisting all 70 000 images of digits

$$\mathcal{P}: \ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.$$

We test these against 500 samples drawn with replacement against one of the sets

$$Q_1 \colon 1, 3, 5, 7, 9,$$
$$Q_2 \colon 0, 1, 3, 5, 7, 9,$$
$$Q_3 \colon 0, 1, 2, 3, 5, 7, 9,$$
$$Q_4 \colon 0, 1, 2, 3, 4, 5, 7, 9,$$
$$Q_5 \colon 0, 1, 2, 3, 4, 5, 6, 7, 9,$$

of respective sizes 35 582, 42 485, 49 475, 56 299 and 63 175. While the samples are in dimension 49, distinguishing images of different digits reduces to a lower-dimensional problem. We consider the Gaussian kernel with the collections of bandwidths $\Lambda(8, 12)$, $\Lambda(10, 14)$ and $\Lambda(12, 16)$ and the Laplace kernel with $\Lambda(10, 14)$, $\Lambda(12, 16)$ and $\Lambda(14, 18)$. As i increases, distinguishing $\mathcal{P}$ from Q_i becomes a more challenging task, which results in a decrease in power.

In the experiments presented in Figure 5 on image data, `MMDAgg`* and `AutoML` achieve the same power and outperform by far all the other tests. This could be due to the fact that the collections for the other aggregated tests consist of very large bandwidths (from $2^8\lambda_{med}$ to $2^{18}\lambda_{med}$). Nonetheless, we observe that `split` and `oracle`, which consider smaller bandwidths, also perform poorly compared to `MMDAgg`* and `AutoML`.

The four aggregated tests with different weighting strategies often match, or even slightly beat, the performance of `oracle` which uses extra data to select an optimal bandwidth. Moreover, they outperform significantly the two adaptive tests `split` and `ost`, as well as the `median` test.

For `MMDAgg`*, using either the Laplace or Gaussian kernel leads to the same performance in this experiment. Furthermore, `MMDAgg`* retains its high power when considering many more kernels as well, as can be seen in the bottom figure of Figure 5 with `MMDAgg`* `All`. Contrary to the previous experiments, we observe that using the Laplace kernel rather than the Gaussian one results in substantially higher power for the four aggregated tests with different weighting strategies. We recall that in the experiments of Figures 3 and 4, we observed that using a Gaussian kernel leads to higher power. This illustrates that the optimal choice of kernel varies depending on the type of data, as such in practice we recommend using `MMDAgg`* `Laplace Gaussian` since it is observed to achieve the highest power obtained by either `MMDAgg`* `Laplace` or `MMDAgg`* `Gaussian` in those three experiments. The test `MMDAgg`* `All` also obtains the same power, but as it considers 12 types of kernels, each with 10 bandwidths, it is computationally more expensive.

The pattern we observed in Figures 3 and 4 of having `MMDAgg increasing` and `MMDAgg decreasing` obtaining the highest and lowest power of our four aggregated tests for the collections of smaller and larger bandwidths, respectively, still holds to some extent in Figure 5 but the differences are less significant. For the collection of bandwidths $\Lambda(12, 16)$ with the Laplace kernel, `MMDAgg centred` does not perform the best, it obtains slightly less power than `MMDAgg uniform` and `MMDAgg increasing` which have almost equal power. Following our interpretation, this simply means that, while $\Lambda(12, 16)$ is an appropriate choice of collection, the optimal bandwidth might be slightly larger than $2^{14}\lambda_{med}$.

Note that, except `MMDAgg`* and `AutoML`, each test obtains similar power when trying to distinguish $\mathcal{P}$ from either Q_3 or Q_4. Recall that Q_3 consists of images of all the digits

except 4, 6 and 8 while Q_4 consists of images of all of them except 6 and 8. One possible explanation could be that these tests distinguish $\mathcal{P}$ from Q_3 mainly by detecting if images of the digit 6 appear in the sample, this would explain why we observe similar power for Q_3 and Q_4, and why the power for Q_5 (consisting of every digit except 8) drops significantly.

We also sometimes observe in Figure 5 that our aggregated tests with different weighting strategies obtain slightly higher power for Q_4 than for Q_3, which might at first seem counter-intuitive. This could be explained by the fact that the optimal bandwidths for distinguishing $\mathcal{P}$ from Q_3 and from Q_4 might be very different, and that the choice of collections of bandwidths presented in Figure 5 are slightly better suited for distinguishing $\mathcal{P}$ from Q_4 than from Q_3. While it is also the case in Figures 3 and 4 that the alternatives with different number of perturbations require different bandwidths to be detected, it looks like in that case considering a collection of five bandwidths which are powers of 2 is enough to adapt to those differences. For the MNIST experiment in Figure 5, it seems that the differences between the optimal bandwidths for Q_3 and Q_4 are more important. Using `MMDAgg`* with its adaptive parameter-free collection of bandwidths solves this problem. An advantage of our aggregated MMDAgg tests is that, even if we fix the collection of bandwidths, they are able to detect differences at various lengthscales, this is not the case for the `median` and `split` tests as those select some specific bandwidth and are only able to detect the differences at the corresponding lengthscale.

5.7 Power experiment: continuous limit of the collection of bandwidths

Our collection of bandwidths for `MMDAgg`* is a discretisation of an interval using $N = 10$ points, which we formally introduced in Section 5.2. As we increase the number of points

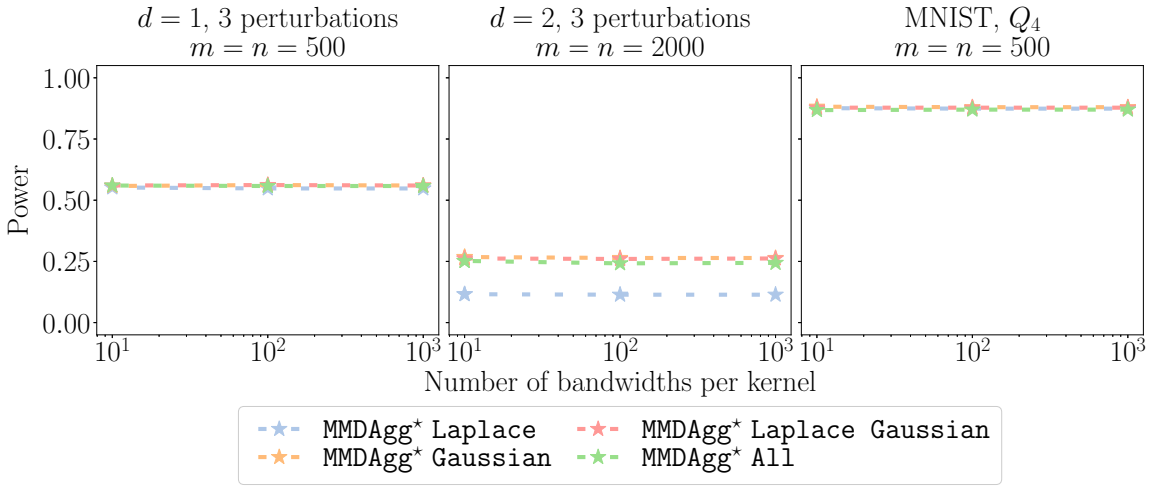

Figure 6: Power experiments varying the size of the collections of bandwidths using perturbed uniform d-dimensional distributions and the MNIST dataset with a wild bootstrap. For `MMDAgg`* `Laplace Gaussian`, two kernels are aggregated, each with a varying number of bandwidths. For `MMDAgg`* `All`, 12 kernels are considered with a varying number of bandwidths (up to 12000 kernels are aggregated).

N, the discretisation becomes finer and in the limit as $N \to \infty$ it corresponds to the whole continuous interval. In Figure 6, we consider the three experiments of Figures 3 to 5 presented in Sections 5.5 and 5.6 for `MMDAgg*` `Gaussian`, `MMDAgg*` `Laplace`, `MMDAgg*` `Laplace Gaussian`, and `MMDAgg*` `All`, with collections of sizes N, N, $2N$ and $12N$, respectively (details in Section 5.4). We vary the number of bandwidths per kernel N to be 10, 100 and 1000. In particular, this means for example that `MMDAgg*` `All` with $N = 1000$ aggregates over $12N = 12000$ kernels. For the one-dimensional uniform setting, we use three perturbations and sample sizes 500. The 2-dimensional case is considered with three perturbations and $m = n = 2000$. For the MNIST experiment, we use Q_4 as an alternative (every digit except 8 and 6) with sample sizes 500.

As in the previous experiments, in Figure 6, the MMDAgg test aggregating both Laplace and Gaussian kernels obtains the highest power achieved by either `MMDAgg*` `Laplace` or `MMDAgg*` `Gaussian`. Considering more types of kernels with `MMDAgg*` `All` does not change the test power. Since it is computationally more expensive, we recommend using `MMDAgg*` `Laplace Gaussian` in practice.

In Figure 6, we observe for all four tests that the test power remains the same when increasing the number of bandwidths per kernel from 10 to 1000. The case $N = 1000$ simulates the continuous limit of the collection of bandwidths, that is, when the full interval is considered without discretisation. First, this shows that our MMDAgg test retains high power even when aggregating up to 12000 kernels, which is only possible due to the way we perform the level correction. Second, this illustrates that the power of the aggregated test with a continuous collection of bandwidths, can also be achieved by the less computationally expensive MMDAgg test with a discretisation of $N = 10$ points.

5.8 Power experiment for image shift detection

In this section, we consider the experiment of Rabanser et al. (2019, Table 1a) on image shift detection on the MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky, 2009) datasets. Ten different types of shifts are applied to either 10%, 50% or 100% of the samples, those include adversarial shifts, class knock-outs, injecting Gaussian noise, and combining rotations/translations/zoom-ins, with different shift strengths (see Section 4 of Rabanser et al. 2019 for details).

We report in Table 1 the test power obtained by our four `MMDAgg*` tests, by our `MMDAgg` `uniform` test with Laplace kernels and with Gaussian kernels, as well as by four versions of the `AutoML` test of Kübler et al. (2022b, Table 1a). We observe that the `MMDAgg*` tests outperform the `MMDAgg` `uniform` tests. For `MMDAgg*`, using the Laplace kernel performs the best, aggregating both Laplace and Gaussian kernels performs better than using only Gaussian kernels and almost as well as using only Laplace kernels, aggregating many more kernels with `MMDAgg*` `All` results in the same power obtained by `MMDAgg*` `Laplace Gaussian`.

Despite using off-the-shelf kernels which are not specifically designed for images, our aggregated `MMDAgg*` tests are still performing within only a few percentage points of state-of-the-art tests based on training models (*e.g.* neural networks) which excel on image data, such as the `AutoML` test.

Table 1: Image shift detection experiment of Rabanser et al. (2019). The numbers reported correspond to the test power averaged over 60 alternatives (each repeated 5 times): 10 different shift types applied on either 10%, 50% or 100% of the image samples drawn from either the MNIST or CIFAR-10 datasets.

Test	Number of samples							
	10	20	50	100	200	500	1000	10000
`MMDAgg`* `Laplace`	0.21	0.29	0.40	0.44	0.47	0.56	0.67	0.83
`MMDAgg`* `Gaussian`	0.19	0.26	0.34	0.42	0.42	0.51	0.62	0.75
`MMDAgg`* `Laplace Gaussian`	0.21	0.27	0.37	0.43	0.45	0.54	0.65	0.80
`MMDAgg`* `All`	0.21	0.27	0.37	0.43	0.46	0.55	0.66	0.80
`MMDAgg uniform (Laplace)`	0.20	0.28	0.40	0.43	0.46	0.52	0.58	0.79
`MMDAgg uniform (Gaussian)`	0.15	0.23	0.33	0.35	0.38	0.44	0.48	0.69
`AutoML (raw)`	0.17	0.24	0.37	0.46	0.50	0.62	0.67	0.87
`AutoML (pre)`	0.18	0.29	0.42	0.47	0.47	0.64	0.65	0.72
`AutoML (class)`	0.19	0.19	0.38	0.46	0.52	0.61	0.67	0.87
`AutoML (bin)`	0.03	0.14	0.31	0.43	0.49	0.51	0.59	0.86

5.9 Overview of additional experimental results

We consider additional experiments in Appendix A, we briefly summarize them here. In Appendix A.1, we verify that all the tests we consider have well-calibrated levels. We then consider widening the collection the bandwidths for our tests `MMDAgg uniform` and `MMDAgg centred` in Appendix A.2, and observe that the associated cost in the power is relatively small. We verify in Appendix A.3 that using a wild bootstrap or permutations results in similar performance; the difference is of non-significant order and is not biased towards one or the other. In Appendix A.4, we show that if one of the sample sizes is fixed to a small number, we cannot obtain high power even if we take the other sample size to be very large. Finally, we increase the sample sizes for the `ost` test in Appendix A.5 and observe that we need extremely large sample sizes to match the performance of `MMDAgg uniform` with 500 or 2000 samples.

6. Conclusion and future work

We have constructed a two-sample hypothesis test, called MMDAgg, which aggregates multiple MMD tests using different kernels/bandwidths. Our test is adaptive over Sobolev balls and does not require data splitting. We have proved that MMDAgg is optimal in the minimax sense over Sobolev balls up to an iterated logarithmic term, for any product of one-dimensional translation invariant characteristic kernels which are absolutely and square integrable. This optimality result also holds under two popular strategies used in estimating the test thresholds, namely the wild bootstrap and permutation procedures. In practice, we propose four weighting strategies which allow the user to incorporate prior knowledge about the collection of bandwidths. We also introduce a parameter-free adaptive collection

of bandwidths, which we recommend using in practice while aggregating over both Gaussian and Laplace kernels, each with multiple bandwidths. This adaptive collection is the discretisation of an interval and in practice we found that using ten bandwidths per kernel performs as well as using the whole continuous interval in the limit. Our MMDAgg test obtains significantly higher power than other state-of-the-art MMD-based two-sample tests in synthetic settings where the smoothness Sobolev assumption is satisfied, this empirically validates our theoretical result of minimax optimality and adaptivity over Sobolev balls. In experiments on image data, we observe that MMDAgg almost matches the power of much more complex two-sample tests relying on training models, such as neural networks, to detect the difference in distributions.

We now discuss three research directions based on this current work, two of which have been explored by Schrab et al. (2022a,b).

First, it would be interesting to consider the two-sample kernel-based test of Jitkrittum et al. (2016), who use adaptive features (in the data space or in the Fourier domain) to construct a linear-time test with good test power. Jitkrittum et al. (2016) require setting aside part of the data to select the kernel bandwidths and the feature locations, by maximizing a proxy for test power. They then perform the test on the remaining data. It would be of interest to develop an approach to learning such adaptive interpretable features without data splitting. Adapting the current results of this work to that setting remain an open challenge.

Second, aggregated tests that are adaptive over Sobolev balls have been constructed for several alternative testing scenarios. The independence testing problem using the Hilbert Schmidt Independence Criterion has been treated by Albert et al. (2022), which is related to the Maximum Mean Discrepancy (Gretton et al., 2012a, Section 7.4). A further setting of interest is goodness-of-fit testing, where a sample is compared against a model. Our theoretical results can directly be applied to goodness-of-fit testing using the MMD, as long as the expectation of the kernel under the model can be computed in closed form. A more challenging problem arises when this expectation cannot be easily computed. In this case, a test may be constructed based on the Kernelised Stein Discrepancy (KSD—Liu et al., 2016; Chwialkowski et al., 2016). This corresponds to computing a Maximum Mean Discrepancy in a modified Reproducing Kernel Hilbert Space, consisting of functions which have zero expectation under the model. Building on the present paper, Schrab et al. (2022a) develop an adaptive aggregated KSDAgg test of goodness-of-fit for the KSD and provide conditions which guarantee high test power for KSDAgg.

Third, our MMDAgg test proposed in this work, the KSDAgg test of Schrab et al. (2022a), and the aggregated HSIC test of Albert et al. (2022), are all quadratic-time hypothesis tests. While quadratic-time tests usually achieve higher power than linear-time tests, this comes at the expense of an important computational cost. To tackle this problem, relying on incomplete U-statistics, Schrab et al. (2022b) propose efficient variants (including linear-time ones) of those three aggregated tests, called MMDAggInc, KSDAggInc and HSICAggInc. They theoretically quantify the cost incurred in the minimax rate over Sobolev balls for this improvement in computational efficiency.

Acknowledgements

We would like to thank the action editor Ingo Steinwart and the anonymous referees for their thorough reviews and suggestions which have helped to significantly improve the paper. Antonin Schrab acknowledges support from the U.K. Research and Innovation under grant number EP/S021566/1. Ilmun Kim acknowledges support from the Yonsei University Research Fund of 2022-22-0289 as well as support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384), and the Korea government (MSIT) RS-2023-00211073. Béatrice Laurent acknowledges the funding by ANITI ANR-19-PI3A-0004. Benjamin Guedj acknowledges partial support by the U.S. Army Research Laboratory and the U.S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1; Benjamin Guedj also acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02. Arthur Gretton acknowledges support from the Gatsby Charitable Foundation.

Overview of Appendices

In Appendix A, we present results of additional experiments. In Appendix B, we explain the relation between using permutations and using the wild bootstrap. We present an efficient implementation of MMDAgg in Appendix C. We highlight the proof strategy of deriving the minimax rate over a Sobolev ball in Appendix D. Finally, we present the proofs of all our results in Appendix E.

Appendix A. Additional experiments

In this section, we verify the level achieved by all the tests considered (Appendix A.1), and run multiple power experiments: widening the collection of bandwidths (Appendix A.2), comparing wild bootstrap and permutations (Appendix A.3), using unbalanced sample sizes (Appendix A.4), and increasing the sample sizes for the `ost` test (Appendix A.5).

A.1 Level experiments

In Table 2, we empirically verify that all the tests we consider have the desired level $\alpha = 0.05$ in the three different settings considered in Figures 3 to 5. For the aggregated MMDAgg tests of Section 5.1, we use the collection of bandwidths $\Lambda(-4, 0)$ for samples drawn from a uniform distribution in one and two dimensions. For samples drawn from the set $\mathcal{P}$ of images of all MNIST digits, we use $\Lambda(10, 14)$ and $\Lambda(12, 16)$ for the Gaussian and Laplace kernels, respectively. To obtain more precise results, we use 5000 repetitions to estimate the levels.

We observe in Table 2 that all the tests have well-calibrated levels, indeed all the estimated levels are relatively close to the prescribed level 0.05. We consider three different types of data and run the tests with the Gaussian and Laplace kernel using either a wild bootstrap or permutations. We note that there is no noticeable trend in the differences in the estimated levels across all those different settings.

Table 2: Level experiments with samples drawn either from d-dimensional uniform distributions or from the MNIST dataset using the Gaussian (G.) and Laplace (L.) kernels with either a wild bootstrap (w.b.) or permutations (p.).

			MMDAgg uniform	MMDAgg centred	MMDAgg increasing	MMDAgg decreasing	median	split	ost
$d = 1$	G.	w.b.	0.0476	0.052	0.0456	0.0434	0.047	0.054	0.0594
		p.	0.0496	0.0532	0.0478	0.0454	0.0468	0.0528	0.0594
	L.	w.b.	0.0474	0.0488	0.0516	0.0504	0.0534	0.05	0.0586
		p.	0.047	0.0482	0.0496	0.0494	0.0522	0.0494	0.0586
$d = 2$	G.	w.b.	0.039	0.0432	0.044	0.0496	0.0464	0.0482	0.0478
		p.	0.0424	0.0446	0.0414	0.0498	0.0466	0.0472	0.0478
	L.	w.b.	0.0382	0.0502	0.0506	0.0478	0.0438	0.0548	0.0502
		p.	0.0418	0.0474	0.0514	0.049	0.0458	0.0548	0.0502
MNIST	G.	w.b.	0.0478	0.0528	0.0474	0.0488	0.0526	0.0498	0.0496
		p.	0.042	0.05	0.0476	0.048	0.055	0.0484	0.0496
	L.	w.b.	0.054	0.052	0.0424	0.0548	0.0518	0.0444	0.05
		p.	0.0526	0.0532	0.0442	0.0554	0.051	0.0448	0.05

		MMDAgg* Laplace	MMDAgg* Gaussian	MMDAgg* Laplace & Gaussian	MMDAgg* All	AutoML
$d = 1$	w.b.	0.0506	0.05	0.0496	0.0492	0.0462
	p.	0.0528	0.0492	0.0496	0.0494	
$d = 2$	w.b.	0.0458	0.0428	0.0434	0.043	0.051
	p.	0.0456	0.0438	0.0438	0.0434	
MNIST	w.b.	0.0624	0.0624	0.0632	0.0594	0.518
	p.	0.0612	0.062	0.0628	0.0622	

A.2 Power experiments: widening the collection of bandwidths

In practice, we might not have strong prior knowledge to guide us in the choice of a collection consisting of only a few bandwidths. For this reason, in practice, we recommend using the adaptive parameter-free collection introduced in Section 5.2 with both Laplace and Gaussian kernels. Nonetheless, it is interesting to study the properties of the family of collections of Section 5.1 consisting of the median bandwidth scaled by powers of 2. In Figure 7, we design an experiment where we start with a collection of 3 bandwidths chosen to be centred around the optimal bandwidth. We then widen the collection of bandwidths and observe how much the power deteriorates as more bandwidths are included in the collection.

We consider collections ranging from 3 to 15 bandwidths for our two tests `MMDAgg uniform` and `MMDAgg centred`, for the three types of data used in Figures 3 to 5. For the perturbed uniform distributions in one and two dimensions, we use the collection of bandwidths $\Lambda(-2-i, -2+i)$ for $i = 1, \ldots, 7$ for both kernels. For the MNIST dataset with

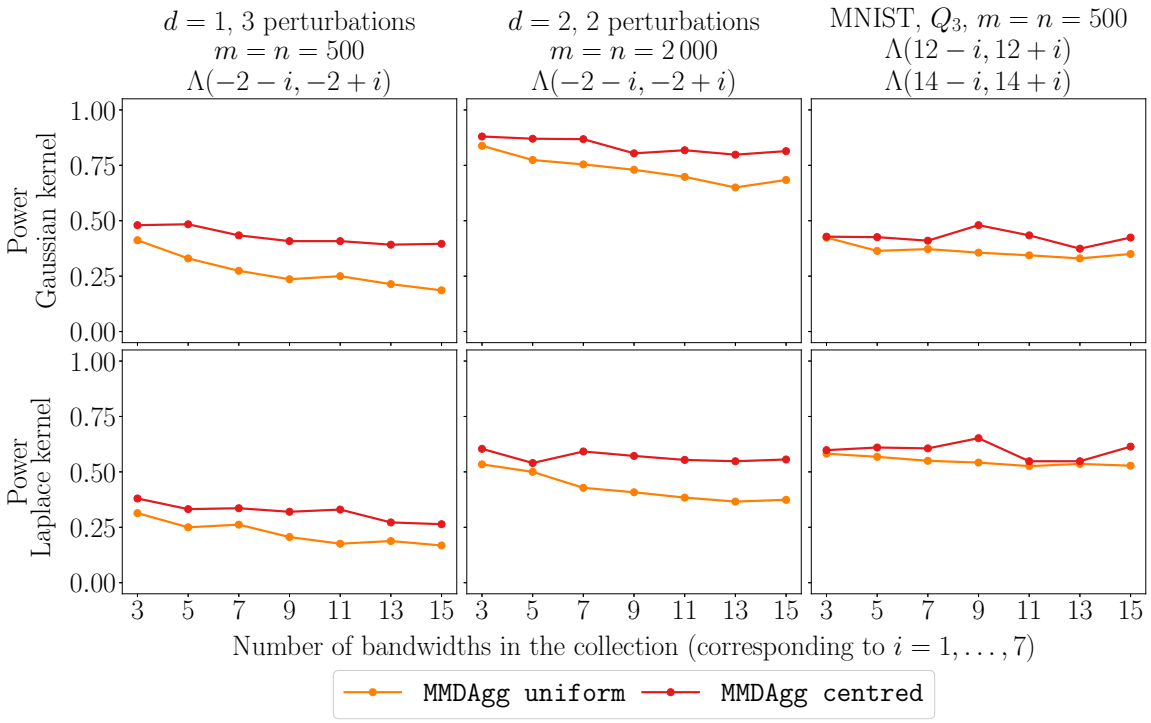

Figure 7: Power experiments varying the size of the collections of bandwidths using perturbed uniform d-dimensional distributions and the MNIST dataset with a wild bootstrap.

the Gaussian and Laplace kernels, we use the collections of bandwidths $\Lambda(12-i, 12+i)$ and $\Lambda(14-i, 14+i)$ for $i = 1, \ldots, 7$, respectively. Those collections are centred as those corresponding to the middle columns of Figures 3 to 5 (for the case $i = 2$), so we expect the bandwidth in the centre of each collection to be a well-calibrated one. For this reason, it makes sense to consider only `MMDAgg uniform` and `MMDAgg centred` in those experiments.

In all the settings considered in Figure 7, we observe only a very small decrease in power when considering a wider collection of bandwidths for `MMDAgg centred`. This is due to the fact that even though we consider more bandwidths, we still put the highest weight on the well-calibrated one in the centre of the collection. Nonetheless, the fact that almost no power is lost when considering more bandwidths for `MMDAgg centred` is a great feature of our test, which is only possible due to the way we perform the level correction for MMDAgg. For the MNIST dataset, we observe a slight increase in power for `MMDAgg centred` with the collections of nine bandwidths for both kernels. This could indicate that, as suggested in Section 5.6, the bandwidths in the centre of the collections are well-calibrated to distinguish $\mathcal{P}$ from Q_4 but are not necessarily the best choice to distinguish $\mathcal{P}$ from Q_3.

Remarkably, the power for `MMDAgg uniform`, which puts equal weights on all the bandwidths, decays relatively slowly and this test does not use the information that the bandwidth in the centre of the collection is a well-calibrated one. So, we expect similar results for any collections of the same sizes which include this bandwidth but not necessarily in the centre of the collection. This means that, in practice, without any prior knowledge, one

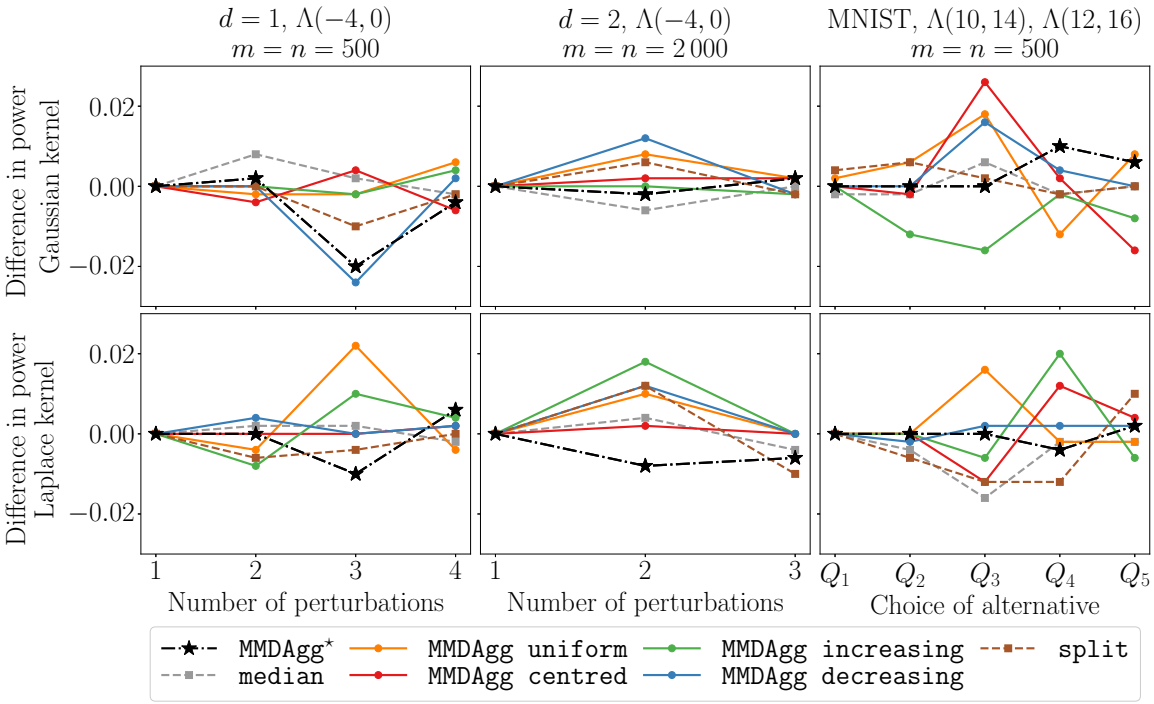

Figure 8: Power experiments considering the difference between using a wild bootstrap or permutations on perturbed uniform d-dimensional distributions and on the MNIST dataset.

can use uniform weights with a relatively wide collection of bandwidths without incurring a considerable loss in power.

A.3 Power experiments: comparing wild bootstrap and permutations

We consider the settings of the experiments presented in Figures 3 to 5 on synthetic and real-world data using the Gaussian and Laplace kernels. We run the same experiments using permutations instead of a wild bootstrap for one collection of bandwidths for each of the different settings. We then consider the power obtained using a wild bootstrap minus the one obtained using permutations, and plot this difference in Figure 8.

The absolute difference in power between using a wild bootstrap or permutations is minimal, it is at most roughly 0.02 and is even considerably smaller in most cases. Furthermore, the difference overall does not seem to be biased towards using either of the two procedures. Since there is no significant difference in power, we suggest using a wild bootstrap when the sample sizes are the same since our implementation of it runs slightly faster in practice. Of course, when the sample sizes are different, one must use permutations.

A.4 Power experiments: using unbalanced sample sizes

In Figure 9, we consider fixing the sample size m and increasing the size n of the other sample, we use permutations since we work with different sample sizes. We consider the

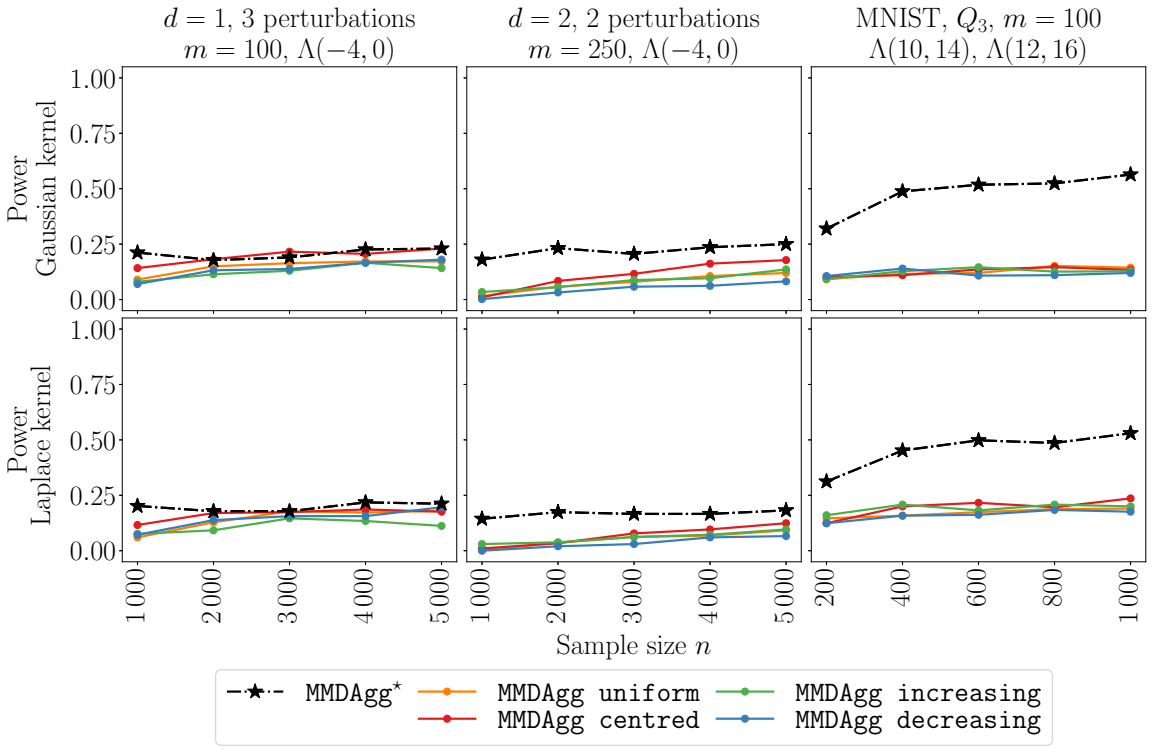

Figure 9: Power experiments with different sample sizes $m \neq n$ on perturbed uniform d-dimensional distributions and on the MNIST dataset using permutations.

settings of Figures 3 to 5 with three and two perturbations for the uniform distributions in one and two dimensions, respectively. For the MNIST dataset, we use the set of images of all digits $\mathcal{P}$ against the set of images Q_3 which does not include the digits 4, 6 and 8.

We observe the same patterns across the six experiments presented in Figure 9. When fixing one of the sample sizes to be small (100 or 250), we cannot achieve power higher than 0.25 (except for `MMDAgg`* on the MNIST data, which as in Figure 5 performs much better than all other tests) by increasing the size of the other sample to be very large (up to 5000). Indeed, we observe that a plateau is reached where considering an even larger sample size does not result in higher power. In some sense, all the information provided by the small sample has already been extracted and using more points for the other sample has almost no effect. As shown in Figures 3 to 5, we can obtain significantly higher power in all of those settings using samples of sizes $m = n = 500$ or $m = n = 2000$. Having access to even more samples overall (5100 instead of 1000) but in such an unbalanced way results in very low power. This shows the importance of having, if possible, balanced datasets with sample sizes of the same order.

A.5 Power experiments: increasing sample sizes for the `ost` test

In Figure 10, we report the results of Figures 3 to 5 for `MMDAgg uniform` and `ost` which are run using the same collections of bandwidths. As previously mentioned, the `ost` test

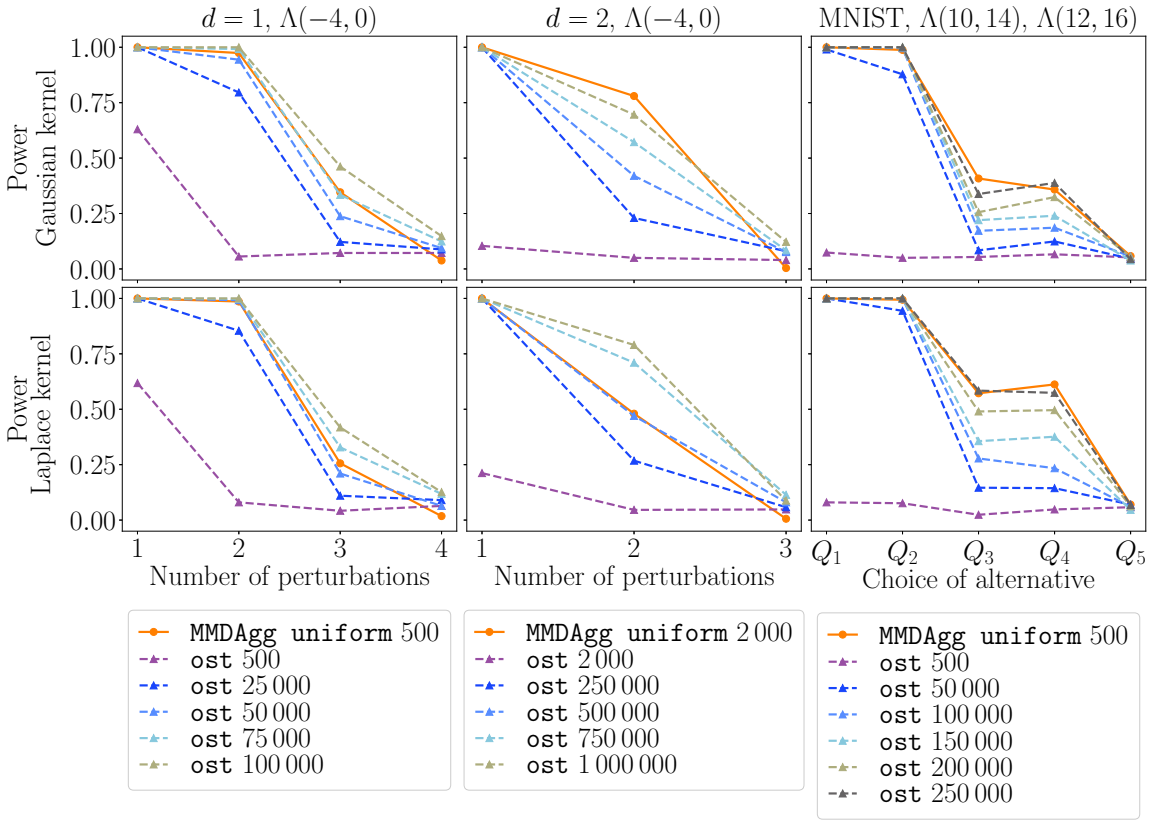

Figure 10: Power experiments increasing the sample sizes for the `ost` test on perturbed uniform d-dimensional distributions and on the MNIST dataset. The legend lists the name of the test followed by the sample sizes used.

is restricted to the use of the linear-time MMD estimate, and hence obtains low power compared to the other tests which use the quadratic-time MMD estimate. We increase the sample sizes for the `ost` test until it matches the power of `MMDAgg uniform` with fixed sample sizes.

For the case of 1-dimensional perturbed uniform distributions, we observe in Figure 10 that `ost` requires sample sizes of 75 000 and 50 000 in order to match the power obtained by `MMDAgg uniform` with $m = n = 500$ for the Gaussian and Laplace kernels, respectively. For the case of 2-dimensional perturbed uniform distributions, one million, and half a million samples are required for the Gaussian and Laplace kernels, respectively, to obtain the same power as `MMDAgg uniform` with 2000 samples. When working with the MNIST dataset, it takes 250 000 samples for `ost` to achieve similar power to the one obtained by `MMDAgg uniform` with 500 samples.

Recall that the MNIST dataset consists of 70 000 images, so it is interesting to see that the power of `ost` keeps increasing for sample sizes which are more than ten times bigger than the size of the dataset. This is due to the use of the linear-time MMD estimate which

for even sample sizes $n = m$ is equal to

$$\frac{2}{n}\sum_{i=1}^{n/2} h_k(X_{2i-1}, X_{2i}, Y_{2i-1}, Y_{2i})$$

for h_k defined as in Equation (5). The two pairs of samples (X_{2i-1}, X_{2i}) and (Y_{2i-1}, Y_{2i}) only appear together as a 4-tuple in this estimate, for $i = 1, \dots, n/2$. So, as long as we do not sample exactly the two same pairs of images together, this creates a new 4-tuple which is considered as new data for this estimate. This explains why considering sample sizes much larger than 70 000 still results in an increase in power.

Appendix B. Relation between permutations and wild bootstrap

In this section, we assume that we have equal sample sizes $m = n$ and show the relation between using permutations and using a wild bootstrap for the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$ defined in Equation (6).

First, we introduce some notation. For a matrix $A = (a_{i,j})_{1\le i,j\le 2n}$, we denote the sum of all its entries by A_+ and denote by A° the matrix A with all the entries

$$\{a_{i,i}, a_{n+i,n+i}, a_{n+i,i}, a_{i,n+i} : i = 1, \dots, n\}$$

set equal to 0. Note that A is composed of four $(n \times n)$-submatrices, and that A° is the matrix A with the diagonal entries of those four submatrices set to 0. We let $\mathbb{1}_n \in \mathbb{R}^{n\times 1}$ denote the vector of length n with all entries equal to 1. We also let $v \coloneqq (\mathbb{1}_n, -\mathbb{1}_n) \in \mathbb{R}^{2n\times 1}$ and note that

$$vv^\top = \begin{pmatrix} \mathbb{1}_n\mathbb{1}_n^\top & -\mathbb{1}_n\mathbb{1}_n^\top \\ -\mathbb{1}_n\mathbb{1}_n^\top & \mathbb{1}_n\mathbb{1}_n^\top \end{pmatrix} \in \mathbb{R}^{2n\times 2n}.$$

We let K_λ denote the kernel matrix $(k_\lambda(U_i, U_j))_{1\le i,j\le 2n}$ where $U_i \coloneqq X_i$ and $U_{n+i} \coloneqq Y_i$ for $i = 1, \dots, n$. We let Tr denote the trace operator and $\circ$ denote the Hadamard product.

By definition of Equation (6), we have

$$\begin{aligned}
\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}(\mathbb{X}_n, \mathbb{Y}_n) &\coloneqq \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} h_\lambda(X_i, X_j, Y_i, Y_j) \\
&= \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} k_\lambda(X_i, X_j) + k_\lambda(Y_i, Y_j) - k_\lambda(X_i, Y_j) - k_\lambda(Y_i, X_j) \\
&= \frac{1}{n(n-1)} \left(K_\lambda^\circ \circ vv^\top\right)_+ \\
&= \frac{1}{n(n-1)} \mathrm{Tr}\left(K_\lambda^\circ vv^\top\right) \\
&= \frac{1}{n(n-1)} \mathrm{Tr}\left(v^\top K_\lambda^\circ v\right) \\
&= \frac{1}{n(n-1)} v^\top K_\lambda^\circ v.
\end{aligned}$$

For the wild bootstrap, as presented in Section 3.2.2, we have n i.i.d. Rademacher random variables $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n)$ with values in $\{-1, 1\}^n$, and let $v_\epsilon \coloneqq (\epsilon, -\epsilon) \in \{-1, 1\}^{2n}$, so that

$$v_\epsilon v_\epsilon^\top = \begin{pmatrix} \epsilon\epsilon^\top & -\epsilon\epsilon^\top \\ -\epsilon\epsilon^\top & \epsilon\epsilon^\top \end{pmatrix} \in \mathbb{R}^{2n \times 2n}.$$

As in Equation (11), we then have

$$\begin{aligned}
\widehat{M}_\lambda^\epsilon &\coloneqq \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \epsilon_i \epsilon_j h_\lambda(X_i, X_j, Y_i, Y_j) \\
&= \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \epsilon_i \epsilon_j k_\lambda(X_i, X_j) + \epsilon_i \epsilon_j k_\lambda(Y_i, Y_j) - \epsilon_i \epsilon_j k_\lambda(X_i, Y_j) - \epsilon_i \epsilon_j k_\lambda(Y_i, X_j) \\
&= \frac{1}{n(n-1)} \left(K_\lambda^\circ \circ v_\epsilon v_\epsilon^\top \right)_+ \\
&= \frac{1}{n(n-1)} \operatorname{Tr}\left(K_\lambda^\circ \, v_\epsilon v_\epsilon^\top \right) \\
&= \frac{1}{n(n-1)} \, v_\epsilon^\top K_\lambda^\circ \, v_\epsilon.
\end{aligned}$$

We introduce more notation. For a matrix $A = (a_{i,j})_{1 \le i,j \le 2n}$ and some permutation $\tau \colon \{1, \dots, 2n\} \to \{1, \dots, 2n\}$, we denote by $A^{\circ\tau}$ the matrix A with all the entries

$$\{a_{\tau(i),\tau(i)}, a_{\tau(n+i),\tau(n+i)}, a_{\tau(n+i),\tau(i)}, a_{\tau(i),\tau(n+i)} : i = 1, \dots, n\}$$

set to be equal to 0. We denote by A_τ the permuted matrix $\left(a_{\tau(i),\tau(j)}\right)_{1 \le i,j \le 2n}$. Similarly, for a vector $w = (w_1, \dots, w_{2n})$, we write the permuted vector as $w_\tau = (w_{\tau(1)}, \dots, w_{\tau(2n)})$.

Recall that $v = (v_1, \dots, v_{2n}) = (\mathbb{1}_n, -\mathbb{1}_n) \in \mathbb{R}^{2n \times 1}$. Similarly to Equation (10) in Section 3.2.1, but for the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}$, given a permutation $\sigma \colon \{1, \dots, 2n\} \to \{1, \dots, 2n\}$, we can define $\widehat{M}_\lambda^\sigma$ as

$$\begin{aligned}
&\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h_\lambda(U_{\sigma(i)}, U_{\sigma(j)}, U_{\sigma(n+i)}, U_{\sigma(n+j)}) \\
&= \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} k_\lambda(U_{\sigma(i)}, U_{\sigma(j)}) + k_\lambda(U_{\sigma(n+i)}, U_{\sigma(n+j)}) - k_\lambda(U_{\sigma(i)}, U_{\sigma(n+j)}) - k_\lambda(U_{\sigma(n+i)}, U_{\sigma(j)}) \\
&= \frac{1}{n(n-1)} \left(\left((K_\lambda)_\sigma \right)^\circ \circ v v^\top \right)_+ \\
&= \frac{1}{n(n-1)} \left((K_\lambda^{\circ\sigma})_\sigma \circ v v^\top \right)_+ \\
&= \frac{1}{n(n-1)} \left(K_\lambda^{\circ\sigma} \circ \left(v v^\top \right)_{\sigma^{-1}} \right)_+ \\
&= \frac{1}{n(n-1)} \operatorname{Tr}\left(K_\lambda^{\circ\sigma} v_{\sigma^{-1}} v_{\sigma^{-1}}^\top \right) \\
&= \frac{1}{n(n-1)} \, v_{\sigma^{-1}}^\top K_\lambda^{\circ\sigma} v_{\sigma^{-1}}.
\end{aligned}$$

Using those formulas, we are able to prove the following proposition, but first we introduce two notions. Fix $\ell \in \{1, \dots, n\}$. We say that a permutation $\tau\colon \{1, \dots, 2n\} \to \{1, \dots, 2n\}$ *fixes* $X_\ell = U_\ell$ and $Y_\ell = U_{n+\ell}$ if $\tau(\ell) = \ell$ and $\tau(n+\ell) = n+\ell$. Moreover, we say that it *swaps* $X_\ell = U_\ell$ and $Y_\ell = U_{n+\ell}$ if $\tau(\ell) = n+\ell$ and $\tau(n+\ell) = \ell$. We denote by $\mathcal{P}$ the set of all permutations which either fix or swap X_i and Y_i for all $i = 1, \dots, n$. Since the identity belongs to $\mathcal{P}$, every element in $\mathcal{P}$ is self-inverse, and the composition of two permutations in $\mathcal{P}$ gives an element in $\mathcal{P}$, the set $\mathcal{P}$ is a subgroup of the permutation group.

Proposition 11 *Assume we have equal sample sizes $m = n$ and that we work with the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}$ defined in Equation (6). Then, using a wild bootstrap is equivalent to using permutations which belong to the subgroup $\mathcal{P}$. There is a one-to-one correspondence between these two procedures.*

Proof First, note that for a permutation $\sigma \in \mathcal{P}$, we either have $\sigma(i) = i$, $\sigma(n+i) = n+i$ or $\sigma(i) = n+i$, $\sigma(n+i) = i$ for $i = 1, \dots, n$. Hence, the two sets

$$\{k_\lambda(U_i, U_i), k_\lambda(U_{n+i}, U_{n+i}), k_\lambda(U_i, U_{n+i}), k_\lambda(U_{n+i}, U_i) : i = 1, \dots, n\}$$

and

$$\big\{k_\lambda(U_{\sigma(i)}, U_{\sigma(i)}), k_\lambda(U_{\sigma(n+i)}, U_{\sigma(n+i)}), k_\lambda(U_{\sigma(i)}, U_{\sigma(n+i)}), k_\lambda(U_{\sigma(n+i)}, U_{\sigma(i)}) : i = 1, \dots, n\big\}$$

are equal, we deduce that for $\sigma \in \mathcal{P}$ we have $K_\lambda^{\circ\sigma} = K_\lambda^\circ$. Moreover, since a permutation $\sigma \in \mathcal{P}$ either fixes or swaps X_i and Y_i for $i = 1, \dots, n$, it must be self-inverse, that is $\sigma^{-1} = \sigma$. For the permutation $\sigma \in \mathcal{P}$, we then have

$$\widehat{M}_\lambda^\sigma = \frac{1}{n(n-1)}\, v_\sigma^\top K_\lambda^\circ v_\sigma.$$

for $v_\sigma = (v_{\sigma(1)}, \dots, v_{\sigma(2n)})$ where $v_i = 1$ and $v_{n+i} = -1$ for $i = 1, \dots, n$. We recall that for the wild bootstrap we have n i.i.d. Rademacher random variables $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n)$ with values in $\{-1, 1\}^n$ and

$$\widehat{M}_\lambda^\epsilon = \frac{1}{n(n-1)}\, v_\epsilon^\top K_\lambda^\circ\, v_\epsilon.$$

where $v_\epsilon \coloneqq (\epsilon, -\epsilon) \in \{-1, 1\}^{2n}$.

We need to show that for a given permutation $\sigma \in \mathcal{P}$ there exists some $\epsilon \in \{-1, 1\}^n$ such that $v_\epsilon = v_\sigma$, and that for a given $\epsilon \in \{-1, 1\}^n$ there exists a permutation $\sigma \in \mathcal{P}$ such that $v_\sigma = v_\epsilon$, and that this correspondence is one-to-one.

Suppose we have a permutation $\sigma \in \mathcal{P}$ and let $\epsilon \coloneqq (v_{\sigma(1)}, \dots, v_{\sigma(n)}) \in \{-1, 1\}^n$. We claim that $v_\sigma = v_\epsilon$, that is, that $(v_{\sigma(1)}, \dots, v_{\sigma(2n)}) = (v_{\sigma(1)}, \dots, v_{\sigma(n)}, -v_{\sigma(1)}, \dots, -v_{\sigma(n)})$, so we need to prove that $v_{\sigma(n+i)} = -v_{\sigma(i)}$ for $i = 1, \dots, n$. As $\sigma \in \mathcal{P}$, for $i = 1, \dots, n$, we either have $\sigma(i) = i$ and $\sigma(n+i) = n+i$ in which case

$$v_{\sigma(n+i)} = v_{n+i} = -1 = -v_i = -v_{\sigma(i)},$$

or $\sigma(i) = n+i$ and $\sigma(n+i) = i$ in which case we have

$$v_{\sigma(n+i)} = v_i = 1 = -v_{n+i} = -v_{\sigma(i)}.$$

This proves the first direction.

Now, suppose we are given $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n)$ i.i.d. Rademacher random variables. We have $v_\epsilon = (\epsilon, -\epsilon) \in \{-1, 1\}^{2n}$ and we need to construct $\sigma \in \mathcal{P}$ such that $v_\sigma = v_\epsilon$, that is, $v_{\sigma(i)} = \epsilon_i$ and $v_{\sigma(n+i)} = -\epsilon_i$ for $i = 1, \dots, n$. We can construct such a permutation $\sigma \in \mathcal{P}$ as follows:

for $i = 1, \dots, n$:
  if $\epsilon_i = 1$ **then let** $\sigma(i) \coloneqq i$ **and** $\sigma(n+i) \coloneqq n+i$ (i.e. σ fixes X_i and Y_i)
    we then have $v_{\sigma(i)} = v_i = 1 = \epsilon_i$ and $v_{\sigma(n+i)} = v_{n+i} = -1 = -\epsilon_i$
  if $\epsilon_i = -1$ **then let** $\sigma(i) \coloneqq n+i$ **and** $\sigma(n+i) \coloneqq i$ (i.e. σ swaps X_i and Y_i)
    we then have $v_{\sigma(i)} = v_{n+i} = -1 = \epsilon_i$ and $v_{\sigma(n+i)} = v_i = 1 = -\epsilon_i$

This proves the second direction.

Our two constructions show that the correspondence is one-to-one. ■

This highlights the relation between those two procedures: using a wild bootstrap is equivalent to using a restricted set of permutations for the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$.

Appendix C. Efficient implementation of MMDAgg (Algorithm 1)

We discuss how to compute Step 1 of Algorithm 1 efficiently. To compute the $|\Lambda|$ kernel matrices for the Gaussian and Laplace kernels, we compute the matrix of pairwise distances only once. For each $\lambda \in \Lambda$, we compute all the values $\left(\widehat{M}^b_{\lambda,1}\right)_{1 \le b \le B_1+1}$ and $\left(\widehat{M}^b_{\lambda,2}\right)_{1 \le b \le B_2}$ together. We compute the sums over the permuted kernel matrices efficiently, in particular we do not want to explicitly permute rows and columns of the kernel matrices as this is computationally expensive.

We start by considering the wild bootstrap case, so we have equal sample sizes $m = n$. We let K_λ denote the kernel matrix $(k_\lambda(U_i, U_j))_{1 \le i,j \le 2n}$ where $U_i \coloneqq X_i$ and $U_{n+i} \coloneqq Y_i$ for $i = 1, \dots, n$. Note that K_λ is composed of four $(n \times n)$-submatrices, we denote by K°_λ the matrix K_λ with the diagonal entries of those four submatrices set to 0. As explained in Appendix B, for n i.i.d. Rademacher random variables $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n)$ with values in $\{-1, 1\}^n$, we have

$$\widehat{M}^\epsilon_\lambda = \frac{1}{n(n-1)} v_\epsilon^\top K^\circ_\lambda v_\epsilon$$

where $v_\epsilon \coloneqq (\epsilon, -\epsilon) \in \{-1, 1\}^{2n}$. We want to extend this to be able to compute $\left(\widehat{M}^{\epsilon^{(b)}}_\lambda\right)_{1 \le b \le B}$ for any $B \in \mathbb{N} \setminus \{0\}$. We can do this by letting R be the $2n \times B$ matrix consisting of stacked vectors $\left(v_{\epsilon^{(b)}}\right)_{1 \le b \le B}$ and computing

$$\frac{1}{n(n-1)} \mathrm{diag}\Big(R^\top K^\circ_\lambda R\Big).$$

Note that, in general, given $2n \times B$ matrices $A = (a_{i,j})_{\substack{1 \le i \le 2n \\ 1 \le j \le B}}$ and $C = (c_{i,j})_{\substack{1 \le i \le 2n \\ 1 \le j \le B}}$, we have

$$\mathrm{diag}\Big(A^\top C\Big) = \mathrm{diag}\left(\left(\sum_{r=1}^{2n} a_{r,i} c_{r,j}\right)_{\substack{1 \le i \le B \\ 1 \le j \le B}}\right) = \left(\sum_{r=1}^{2n} a_{r,i} c_{r,i}\right)_{1 \le i \le B} =: \sum_{\text{rows}} A \circ C$$

where $\circ$ denotes the Hadamard product and where $\sum_{\text{rows}}$ takes a matrix as input and outputs a vector which is the sum of the row vectors of the matrix. We deduce that

$$\operatorname{diag}\Big(R^\top K_\lambda^\circ R\Big) = \sum_{\text{rows}} R \circ (K_\lambda^\circ R).$$

We found that this way of computing the values $\big(\widehat{M}_\lambda^{\epsilon^{(b)}}\big)_{1\le b\le B}$ is computationally faster than other alternatives. Letting $\mathbb{1}_n$ denote the vector of length n with all entries equal to 1, we can obtain an efficient version for Step 1 of Algorithm 1 using a wild bootstrap as follows.

Efficient Step 1 of Algorithm 1 using a wild bootstrap:
generate $n \times (B_1 + B_2 + 1)$ matrix $\widetilde{R}$ of Rademacher random variables
concatenate $\tilde{R}$ and $-\tilde{R}$ to form the $2n \times (B_1 + B_2 + 1)$ matrix R
replace the $(B_1+1)^{\text{th}}$ column of R with the vector $(\mathbb{1}_n, -\mathbb{1}_n)$
for $\lambda \in \Lambda$:
    compute kernel matrix K_λ° with zero diagonals for its four submatrices
    compute $\frac{1}{n(n-1)} \sum_{\text{rows}} R \circ (K_\lambda^\circ R)$ to get $\big(\widehat{M}_{\lambda,1}^1, \ldots, \widehat{M}_{\lambda,1}^{B_1+1}, \widehat{M}_{\lambda,2}^1, \ldots, \widehat{M}_{\lambda,2}^{B_2}\big)$
    $\big(\widehat{M}_{\lambda,1}^{\bullet 1}, \ldots, \widehat{M}_{\lambda,1}^{\bullet B_1+1}\big) = \texttt{sort\_by\_ascending\_order}\big(\widehat{M}_{\lambda,1}^1, \ldots, \widehat{M}_{\lambda,1}^{B_1+1}\big)$

Before tackling the case of permutations, we recall the main steps of the strategy used for the wild bootstrap case. As explained in Appendix B, we first noted that

$$\begin{aligned}
\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}(\mathbb{X}_n, \mathbb{Y}_n) &= \left(K_\lambda^\circ \circ \begin{pmatrix} \mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) & -\mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) \\ -\mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) & \mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) \end{pmatrix} \right)_+ \\
&= \frac{1}{n(n-1)} \left(K_\lambda^\circ \circ vv^\top \right)_+ \\
&= \frac{1}{n(n-1)}\, v^\top K_\lambda^\circ\, v.
\end{aligned}$$

for $v \coloneqq (\mathbb{1}_n, -\mathbb{1}_n) \in \mathbb{R}^{2n\times 1}$, where A_+ denotes the sum all the entries of a matrix A. We then observed that it was enough to replace the vector v with $v_\epsilon \coloneqq (\epsilon, -\epsilon) \in \{-1, 1\}^{2n}$ to obtain

$$\widehat{M}_\lambda^\epsilon = \frac{1}{n(n-1)}\, v_\epsilon^\top K_\lambda^\circ\, v_\epsilon.$$

The whole reasoning was based on the fact that we could rewrite the matrix

$$\begin{pmatrix} \mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) & -\mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) \\ -\mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) & \mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) \end{pmatrix}$$

as an outer product of vectors.

Now, we consider the permutation-based procedure. For a square matrix A, we let A^0 denote the matrix A with its diagonal entries set equal to 0. We have

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}_m,\mathbb{Y}_n) &= \frac{1}{m(m-1)}\sum_{1\le i\ne i'\le m} k(X_i,X_{i'}) + \frac{1}{n(n-1)}\sum_{1\le j\ne j'\le n} k(Y_j,Y_{j'}) \\ &\qquad - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(X_i,Y_j) \\ &= \left(K^0_\lambda \circ \begin{pmatrix} \mathbb{1}_m\mathbb{1}_m^\top/(m(m-1)) & -\mathbb{1}_m\mathbb{1}_n^\top/(mn) \\ -\mathbb{1}_n\mathbb{1}_m^\top/(mn) & \mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) \end{pmatrix}\right)_+\end{aligned}$$

where it is not possible to rewrite the matrix as an outer product of vectors. Instead, we break it down into a sum of three outer products of vectors.

$$\begin{aligned}\begin{pmatrix} \mathbb{1}_m\mathbb{1}_m^\top/(m(m-1)) & -\mathbb{1}_m\mathbb{1}_n^\top/(mn) \\ -\mathbb{1}_n\mathbb{1}_m^\top/(mn) & \mathbb{1}_n\mathbb{1}_n^\top/(n(n-1)) \end{pmatrix} &= \left(\frac{1}{m(m-1)} - \frac{1}{mn}\right)\begin{pmatrix} \mathbb{1}_m\mathbb{1}_m^\top & \mathbb{0}_m\mathbb{0}_n^\top \\ \mathbb{0}_n\mathbb{0}_m^\top & \mathbb{0}_n\mathbb{0}_n^\top \end{pmatrix} \\ &\quad + \left(\frac{1}{n(n-1)} - \frac{1}{mn}\right)\begin{pmatrix} \mathbb{0}_m\mathbb{0}_m^\top & \mathbb{0}_m\mathbb{0}_n^\top \\ \mathbb{0}_n\mathbb{0}_m^\top & \mathbb{1}_n\mathbb{1}_n^\top \end{pmatrix} \\ &\quad + \frac{1}{mn}\begin{pmatrix} \mathbb{1}_m\mathbb{1}_m^\top & -\mathbb{1}_m\mathbb{1}_n^\top \\ -\mathbb{1}_n\mathbb{1}_m^\top & \mathbb{1}_n\mathbb{1}_n^\top \end{pmatrix} \\ &= \frac{n-m+1}{mn(m-1)}uu^\top + \frac{m-n+1}{mn(n-1)}ww^\top + \frac{1}{mn}vv^\top\end{aligned}$$

where $u \coloneqq (\mathbb{1}_m, \mathbb{0}_n)$, $w \coloneqq (\mathbb{0}_m, -\mathbb{1}_n)$ and $v \coloneqq (\mathbb{1}_m, -\mathbb{1}_n)$ of shapes $(m+n)\times 1$, with $\mathbb{0}_n$ denoting the vector of length n with all entries equal to 0. Using this fact, we obtain that

$$\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}_m,\mathbb{Y}_n) = \frac{n-m+1}{mn(m-1)}u^\top K^0_\lambda u + \frac{m-n+1}{mn(n-1)}w^\top K^0_\lambda w + \frac{1}{mn}v^\top K^0_\lambda v.$$

For a vector $a = (a_1,\dots,a_\ell)$ and a permutation $\tau\colon \{1,\dots,\ell\} \to \{1,\dots,\ell\}$, we denote the permuted vector as $a_\tau = (a_{\tau(1)},\dots,a_{\tau(\ell)})$. Recall that $U_i \coloneqq X_i$, $i = 1,\dots,m$ and $U_{m+j} \coloneqq Y_j$, $j = 1,\dots,n$. Consider a permutation $\sigma\colon \{1,\dots,m+n\} \to \{1,\dots,m+n\}$ and let $\mathbb{X}^\sigma_m \coloneqq \big(U_{\sigma(i)}\big)_{1\le i\le m}$ and $\mathbb{Y}^\sigma_n \coloneqq \big(U_{\sigma(m+j)}\big)_{1\le j\le n}$. Following a similar reasoning to the one presented in Appendix B, we find

$$\begin{aligned}\widehat{M}^\sigma_\lambda &\coloneqq \widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}^\sigma_m,\mathbb{Y}^\sigma_n) \\ &= \frac{n-m+1}{mn(m-1)}u_{\sigma^{-1}}^\top K^0_\lambda u_{\sigma^{-1}} + \frac{m-n+1}{mn(n-1)}w_{\sigma^{-1}}^\top K^0_\lambda w_{\sigma^{-1}} + \frac{1}{mn}v_{\sigma^{-1}}^\top K^0_\lambda v_{\sigma^{-1}}\end{aligned}$$

since $\big\{k_\lambda(X_i,Y_i) : i = 1,\dots,n\big\} = \big\{k_\lambda(X_{\sigma(i)},Y_{\sigma(i)}) : i = 1,\dots,n\big\}$. The aim is to compute $\big(\widehat{M}^{\sigma^{(b)}}_\lambda\big)_{1\le b\le B}$ efficiently for any $B \in \mathbb{N}\setminus\{0\}$. We let U, V and W denote the $(m+n)\times B$ matrices of stacked vectors $\big(u_{\sigma^{(b)^{-1}}}\big)_{1\le b\le B}$, $\big(v_{\sigma^{(b)^{-1}}}\big)_{1\le b\le B}$ and $\big(w_{\sigma^{(b)^{-1}}}\big)_{1\le b\le B}$, respectively. We are then able to compute $\big(\widehat{M}^{\sigma^{(b)}}_\lambda\big)_{1\le b\le B}$ as

$$\begin{aligned}&\frac{n-m+1}{mn(m-1)}\mathrm{diag}\Big(U^\top K^0_\lambda U\Big) + \frac{m-n+1}{mn(n-1)}\mathrm{diag}\Big(W^\top K^0_\lambda W\Big) + \frac{1}{mn}\mathrm{diag}\Big(V^\top K^0_\lambda V\Big) \\ &= \frac{n-m+1}{mn(m-1)}\sum_{\text{rows}} U\circ\big(K^0_\lambda U\big) + \frac{m-n+1}{mn(n-1)}\sum_{\text{rows}} W\circ\big(K^0_\lambda W\big) + \frac{1}{mn}\sum_{\text{rows}} V\circ\big(K^0_\lambda V\big).\end{aligned}$$

Since the inverse map for permutations is a bijection between the space of all permutations and itself, it follows that uniformly generating B permutations and taking their inverses is equivalent to directly uniformly generating B permutations. So, in practice, we can simply uniformly generate permutations $\tau^{(1)}, \dots, \tau^{(B)}$ and assume that these correspond to ${\sigma^{(1)}}^{-1}, \dots, {\sigma^{(B)}}^{-1}$ for uniformly generated permutations $\sigma^{(1)}, \dots, \sigma^{(B)}$. We can now present an efficient version for Step 1 of Algorithm 1 using permutations.

Efficient Step 1 using permutations:
construct $(m+n) \times (B_1 + B_2 + 1)$ matrix U of stacked vectors of $(\mathbb{1}_m, \mathbb{0}_n)$
construct $(m+n) \times (B_1 + B_2 + 1)$ matrix V of stacked vectors of $(\mathbb{1}_m, -\mathbb{1}_n)$
construct $(m+n) \times (B_1 + B_2 + 1)$ matrix W of stacked vectors of $(\mathbb{0}_m, -\mathbb{1}_n)$
use $B_1 + B_2$ permutations to permute the elements of the columns of U, V and W without permuting the elements of the $(B_1+1)^{\text{th}}$ columns of U, V and W
for $\lambda \in \Lambda$:
compute kernel matrix K_λ^0 with zero diagonals
compute $\big(\widehat{M}_{\lambda,1}^1, \dots, \widehat{M}_{\lambda,1}^{B_1+1}, \widehat{M}_{\lambda,2}^1, \dots, \widehat{M}_{\lambda,2}^{B_2}\big)$ as

$$\frac{n-m+1}{mn(m-1)} \sum_{\text{rows}} U \circ \big(K_\lambda^0 U\big) + \frac{m-n+1}{mn(n-1)} \sum_{\text{rows}} W \circ \big(K_\lambda^0 W\big) + \frac{1}{mn} \sum_{\text{rows}} V \circ \big(K_\lambda^0 V\big)$$

$$\big(\widehat{M}_{\lambda,1}^{\bullet 1}, \dots, \widehat{M}_{\lambda,1}^{\bullet B_1+1}\big) = \texttt{sort\_by\_ascending\_order}\big(\widehat{M}_{\lambda,1}^1, \dots, \widehat{M}_{\lambda,1}^{B_1+1}\big)$$

Appendix D. Lower bound on the minimax rate over a Sobolev ball

We claim that the two-sample minimax rate of testing over the Sobolev ball $\mathcal{S}_d^s(R)$ is $n^{-2s/(4s+d)}$. Formally, let $\alpha, \beta \in (0,1)$, $d \in \mathbb{N} \setminus \{0\}$ and $M, s, R \in (0, \infty)$, we claim that there exists some positive constant $C_0'(M, d, s, R, \alpha, \beta)$ such that

$$\underline{\rho}(\mathcal{S}_d^s(R), \alpha, \beta, M) \coloneqq \inf_{\Delta_\alpha} \rho(\Delta_\alpha, \mathcal{S}_d^s(R), \beta, M) \geq C_0'(M, d, s, R, \alpha, \beta)\, n^{-2s/(4s+d)} \tag{18}$$

where the infimum is taken over all tests Δ_α of non-asymptotic level α and where $c' \leq \frac{m}{n} \leq C'$ for some positive constants c' and C'.

The proof mirrors the reasoning of Albert et al. (2022, Section 4, Theorem 4) who derive the independence minimax rate of testing $n^{-2s/(4s+d_1+d_2)}$ over the Sobolev ball $\mathcal{S}_{d_1+d_2}^s(R)$ considering paired samples on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. While the case $d = 1$ is not covered by their framework, the proof extends naturally to it. We illustrate the main reasoning behind the proof as this is of interest for our experiments in Section 5.5.

First, note that the minimax rate $\underline{\rho}(\mathcal{S}_d^s(R), \alpha, \beta, M)$ is

$$\begin{aligned}
\inf_{\Delta_\alpha} \rho(\Delta_\alpha, \mathcal{S}_d^s(R), \beta, M) &= \inf_{\Delta_\alpha} \inf \left\{ \tilde{\rho} > 0 : \sup_{(p,q) \in \mathcal{F}_{\tilde{\rho}}^M(\mathcal{S}_d^s(R))} \mathbb{P}_{p \times q}(\Delta_\alpha(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta \right\} \\
&= \inf \left\{ \tilde{\rho} > 0 : \inf_{\Delta_\alpha} \sup_{(p,q) \in \mathcal{F}_{\tilde{\rho}}^M(\mathcal{S}_d^s(R))} \mathbb{P}_{p \times q}(\Delta_\alpha(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta \right\}
\end{aligned}$$

where $\mathcal{F}_{\tilde{\rho}}^{M}(\mathcal{S}_d^s(R)) \coloneqq \{(p,q) : \max(\|p\|_\infty, \|q\|_\infty) \leq M, p-q \in \mathcal{S}_d^s(R),\ \|p-q\|_2 > \tilde{\rho}\}$. Hence, to prove Equation (18) it suffices to construct two probability densities p and q on $\mathbb{R}^d$ which satisfy $\max(\|p\|_\infty, \|q\|_\infty) \leq M,\ p-q \in \mathcal{S}_d^s(R),\ \|p-q\|_2 < C_1'(m+n)^{-2s/(4s+d)}$ for some $C_1' > 0$, and for which $\mathbb{P}_{p\times q}(\Delta_\alpha'(\mathbb{X}_m, \mathbb{Y}_n) = 0) > \beta$ holds for all tests Δ_α with non-asymptotic level α. Intuitively, one constructs densities which are close enough in L^2-norm so that any test with non-asymptotic level α fails to distinguish them from one another.

Albert et al. (2022, Section 4) show that a suitable choice of p and q is to take the uniform probability density on $[0,1]^d$ and a perturbed version of it. To construct the latter, first define for all $u \in \mathbb{R}$ the function

$$G(u) \coloneqq \exp\left(-\frac{1}{1-(4u+3)^2}\right)\mathbb{1}_{(-1,-\frac{1}{2})}(u) - \exp\left(-\frac{1}{1-(4u+1)^2}\right)\mathbb{1}_{(-\frac{1}{2},0)}(u)$$

which is plotted in Figure 2 in Section 5.5. Consider $P \in \mathbb{N}\setminus\{0\}$ and some vector $\theta = (\theta_\nu)_{\nu\in\{1,\dots,P\}^d} \in \{-1,1\}^{P^d}$ of length P^d with entries either -1 or 1 which is indexed by the P^d d-dimensional elements of $\{1,\dots,P\}^d$. Then, the perturbed uniform density is defined as

$$f_\theta(u) \coloneqq \mathbb{1}_{[0,1]^d}(u) + P^{-s}\sum_{\nu\in\{1,\dots,P\}^d}\theta_\nu\prod_{i=1}^{d}G\left(Pu_i-\nu_i\right) \tag{19}$$

for all $u \in \mathbb{R}^d$. As illustrated in Figure 2, this indeed corresponds to a uniform probability density with P perturbations along each dimension.

This construction by Albert et al. (2022) is a generalisation of the detailed construction of Butucea (2007, Section 5) for the 1-dimensional case for goodness-of-fit testing. We also point out the work of Li and Yuan (2019, Theorems 5, part (ii)) who use a similar construction to show that, for any alternative[10] in $\mathcal{F}_{\tilde{\rho}}^{M}(\mathcal{S}_d^s(R))$ with rate $\tilde{\rho}$ smaller or equal to $n^{-2s/(4s+d)}$, there exists some $\alpha \in (0,1)$ such that any test with asymptotic level α must have power asymptotically strictly smaller than one. The original idea behind all those constructions is due to Ingster (1987, 1993b) with his work on nonparametric minimax rates for goodness-of-fit testing.

Appendix E. Proofs

In this section, we prove the statements presented in Section 3. We first introduce some standard results.

First, recall that we assume $m \leq n$ and $n \leq Cm$ for some constant $C \geq 1$ as in Equation (7). It follows that

$$\frac{1}{m}+\frac{1}{n} \leq \frac{C+1}{n} = \frac{2(C+1)}{2n} \leq \frac{2(C+1)}{m+n} \quad \text{and} \quad \frac{1}{m+n} \leq \frac{2}{m+n} \leq \frac{1}{m}+\frac{1}{n}. \tag{20}$$

10. To be more precise, the alternatives considered by Li and Yuan (2019) require both p and q to belong to the Sobolev ball (rather than just $p-q$ in our case) but do not require these densities to be bounded. Moreover, they use the non-homogeneous Sobolev space definition (with an extra '+1' in the weighting term) while we use the homogeneous definition (without that extra term).

For the kernels $K_1, \dots, K_d$ satisfying the properties presented in Section 3.1, we define the constants

$$\kappa_1(d) \coloneqq \prod_{i=1}^{d} \int_{\mathbb{R}} |K_i(x_i)| \mathrm{d}x_i < \infty \quad \text{and} \quad \kappa_2(d) \coloneqq \prod_{i=1}^{d} \int_{\mathbb{R}} K_i(x_i)^2 \, \mathrm{d}x_i < \infty \tag{21}$$

which are well-defined as $K_i \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ for $i = 1, \dots, d$ by assumption. We do not make explicit the dependence on $K_1, \dots, K_d$ in the constants as we consider those to be chosen *a priori*. Moreover, we often use the kernel properties of k_λ presented in Equation (8), that are

$$\int_{\mathbb{R}^d} k_\lambda(x, y) \mathrm{d}x = \prod_{i=1}^{d} \frac{1}{\lambda_i} \int_{\mathbb{R}} K_i\Big(\frac{x_i - y_i}{\lambda_i}\Big) \mathrm{d}x_i = \prod_{i=1}^{d} \int_{\mathbb{R}} K_i(x_i') \mathrm{d}x_i' = 1$$

and

$$\int_{\mathbb{R}^d} k_\lambda(x, y)^2 \mathrm{d}x = \prod_{i=1}^{d} \frac{1}{\lambda_i^2} \int_{\mathbb{R}} K_i\Big(\frac{x_i - y_i}{\lambda_i}\Big)^2 \mathrm{d}x_i = \frac{1}{\lambda_1 \cdots \lambda_d} \prod_{i=1}^{d} \int_{\mathbb{R}} K_i\big(x_i'\big)^2 \, \mathrm{d}x_i' = \frac{\kappa_2}{\lambda_1 \cdots \lambda_d}.$$

We often use in our proofs the standard result that, for $a_1, \dots, a_\ell \in \mathbb{R}$, we have

$$\Big(\sum_{i=1}^{\ell} a_i\Big)^2 \leq \Big(\sum_{i=1}^{\ell} 1^2\Big)\Big(\sum_{i=1}^{\ell} a_i^2\Big) = \ell \sum_{i=1}^{\ell} a_i^2$$

which holds by Cauchy–Schwarz inequality.

In our proofs, we show that there exist some constants which are large enough so that our results hold. We keep track of those constants and show how they depend on each other. The aim is to show that such constants exist, we do not focus on obtaining the tightest constants possible.

E.1 Proof of Proposition 1

Recall that in Sections 3.2.1 and 3.2.2 we have constructed elements $\big(\widehat{M}_\lambda^b\big)_{1 \leq b \leq B+1}$ for the two MMD estimators defined in Equations (3) and (6), respectively. The first one uses permutations while the second uses a wild bootstrap. For those two cases, we first show that the elements $\big(\widehat{M}_\lambda^b\big)_{1 \leq b \leq B+1}$ are exchangeable under the null hypothesis $\mathcal{H}_0 \colon p = q$. We are then able to prove that the test $\Delta_\alpha^{\lambda, B}$ has the prescribed level using the exchangeability of $\big(\widehat{M}_\lambda^b\big)_{1 \leq b \leq B+1}$.

Exchangeability using permutations as in Section 3.2.1.

Recall that in this case we have permutations $\sigma^{(1)}, \dots, \sigma^{(B)}$ of $\{1, \dots, m+n\}$. We also have $\widehat{M}_\lambda^b \coloneqq \widehat{\mathrm{MMD}}^2_{\lambda, \mathtt{a}}(\mathbb{X}_m^{\sigma^{(b)}}, \mathbb{Y}_n^{\sigma^{(b)}})$ for $b = 1, \dots, B$ and $\widehat{M}_\lambda^{B+1} \coloneqq \widehat{\mathrm{MMD}}^2_{\lambda, \mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n)$. Following the same reasoning as in the proof of Albert et al. (2022, Proposition 1, Equation C.1), we can deduce that $\big(\widehat{M}_\lambda^b\big)_{1 \leq b \leq B+1}$ are exchangeable under the null hypothesis. The only difference is that they work with the Hilbert Schmidt Independence Criterion rather than with the Maximum Mean Discrepancy, but this does not affect the reasoning of the proof.

Exchangeability using a wild bootstrap as in Section 3.2.2.

The exchangeability under the null using the wild bootstrap follows from the one using permutations since by Proposition 11 using the wild bootstrap corresponds to using a subgroup of the permutation group. We show below how the original statistic and the permuted one can be seen to have the same distribution under the null.

For $b = 1, \dots, B$, we have n i.i.d. Rademacher random variables $\epsilon^{(b)} \coloneqq \left(\epsilon_1^{(b)}, \dots, \epsilon_n^{(b)}\right)$ with values in $\{-1, 1\}^n$ and

$$\widehat{M}_\lambda^b \coloneqq \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \epsilon_i^{(b)} \epsilon_j^{(b)} h_\lambda(X_i, X_j, Y_i, Y_j)$$

where h_λ is defined in Equation (5). We also have

$$\widehat{M}_\lambda^{B+1} \coloneqq \widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}(\mathbb{X}_n, \mathbb{Y}_n) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_\lambda(X_i, X_j, Y_i, Y_j).$$

By the reproducing property of the kernel k_λ, we have

$$\begin{aligned}
\left(\sup_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n \big(f(X_i) - f(Y_i)\big) \right)^2 &= \left(\sup_{f \in \mathcal{F}_\lambda} \left\langle f, \frac{1}{n} \sum_{i=1}^n k_\lambda(X_i, \cdot) - \frac{1}{n} \sum_{i=1}^n k_\lambda(Y_i, \cdot) \right\rangle_{\mathcal{H}_{k_\lambda}} \right)^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n k_\lambda(X_i, \cdot) - \frac{1}{n} \sum_{i=1}^n k_\lambda(Y_i, \cdot) \right\|^2_{\mathcal{H}_{k_\lambda}} \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_\lambda(X_i, X_j, Y_i, Y_j)
\end{aligned}$$

where $\mathcal{F}_\lambda \coloneqq \{f \in \mathcal{H}_{k_\lambda} : \|f\|_{\mathcal{H}_{k_\lambda}} \leq 1\}$. Under the null hypothesis $\mathcal{H}_0 \colon p = q$, all the samples $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ are independent and identically distributed. So, the distribution of $\left(\sup_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n \big(f(X_i) - f(Y_i)\big)\right)^2$ does not change if we randomly exchange X_i and Y_i for each $i = 1, \dots, n$. This can be formalized using n i.i.d. Rademacher random variables $\epsilon_1, \dots, \epsilon_n$, we have

$$\left(\sup_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n \big(f(X_i) - f(Y_i)\big) \right)^2 \overset{d}{=}_{\mathcal{H}_0} \left(\sup_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(f(X_i) - f(Y_i)\big) \right)^2,$$

where the notation $\overset{d}{=}_{\mathcal{H}_0}$ means that the two random variables have the same distribution under the null hypothesis $\mathcal{H}_0 \colon p = q$. Since we also have

$$\begin{aligned}
\left(\sup_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(f(X_i) - f(Y_i)\big) \right)^2 &= \left(\sup_{f \in \mathcal{F}_\lambda} \left\langle f, \frac{1}{n} \sum_{i=1}^n \epsilon_i k_\lambda(X_i, \cdot) - \frac{1}{n} \sum_{i=1}^n \epsilon_i k_\lambda(Y_i, \cdot) \right\rangle_{\mathcal{H}_{k_\lambda}} \right)^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i k_\lambda(X_i, \cdot) - \frac{1}{n} \sum_{i=1}^n \epsilon_i k_\lambda(Y_i, \cdot) \right\|^2_{\mathcal{H}_{k_\lambda}} \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \epsilon_i \epsilon_j h_\lambda(X_i, X_j, Y_i, Y_j),
\end{aligned}$$

we obtain

$$\sum_{i=1}^{n}\sum_{j=1}^{n} h_\lambda(X_i, X_j, Y_i, Y_j) \underset{\mathcal{H}_0}{\overset{d}{=}} \sum_{i=1}^{n}\sum_{j=1}^{n} \epsilon_i\epsilon_j h_\lambda(X_i, X_j, Y_i, Y_j).$$

Since $\epsilon_i^2 = 1$ for $i = 1, \dots, n$, subtracting $\sum_{i=1}^{n} h_\lambda(X_i, X_i, Y_i, Y_i)$ from both sides, we get

$$\sum_{1\le i\neq j\le n} h_\lambda(X_i, X_j, Y_i, Y_j) \underset{\mathcal{H}_0}{\overset{d}{=}} \sum_{1\le i\neq j\le n} \epsilon_i\epsilon_j h_\lambda(X_i, X_j, Y_i, Y_j).$$

Level of the test.

Following a similar reasoning to the one presented by Albert et al. (2022, Proposition 1), we have

$$\begin{aligned}
\Delta_\alpha^{\lambda,B}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_B) = 1 \quad &\Longleftrightarrow \quad \widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B \big| \mathbb{X}_m, \mathbb{Y}_n\big) \\
&\Longleftrightarrow \quad \widehat{M}_\lambda^{B+1} > \widehat{M}_\lambda^{\bullet\lceil (B+1)(1-\alpha)\rceil} \\
&\Longleftrightarrow \quad \sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}_\lambda^b < \widehat{M}_\lambda^{B+1}\Big) \ge \lceil (B+1)(1-\alpha)\rceil \\
&\Longleftrightarrow \quad B+1-\sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}_\lambda^b < \widehat{M}_\lambda^{B+1}\Big) \le B+1-\lceil (B+1)(1-\alpha)\rceil \\
&\Longleftrightarrow \quad \sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}_\lambda^b \ge \widehat{M}_\lambda^{B+1}\Big) \le \lfloor \alpha(B+1)\rfloor \\
&\Longleftrightarrow \quad \sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}_\lambda^b \ge \widehat{M}_\lambda^{B+1}\Big) \le \alpha(B+1) \\
&\Longleftrightarrow \quad \frac{1}{B+1}\left(1+\sum_{b=1}^{B} \mathbb{1}\Big(\widehat{M}_\lambda^b \ge \widehat{M}_\lambda^{B+1}\Big)\right) \le \alpha
\end{aligned}$$

where we have used the fact that $B+1-\lceil (B+1)(1-\alpha)\rceil = \lfloor \alpha(B+1)\rfloor$. Using the exchangeability of $\big(\widehat{M}_\lambda^b\big)_{1\le b\le B+1}$, the result of Romano and Wolf (2005a, Lemma 1) guarantees that

$$\mathbb{P}_{p\times p\times r}\left(\frac{1}{B+1}\left(1+\sum_{b=1}^{B} \mathbb{1}\Big(\widehat{M}_\lambda^b \ge \widehat{M}_\lambda^{B+1}\Big)\right) \le \alpha\right) \le \ \alpha.$$

We deduce that

$$\mathbb{P}_{p\times p\times r}\Big(\Delta_\alpha^{\lambda,B}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_B) = 1\Big) \le \alpha.$$

E.2 Proof of Lemma 2

Let

$$\mathcal{A} \coloneqq \Big\{\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) \le \widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B \big| \mathbb{X}_m, \mathbb{Y}_n\big)\Big\}$$

and

$$\mathcal{B} \coloneqq \left\{ \mathrm{MMD}^2_\lambda(p,q) \geq \sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big)} + \widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B \big| \mathbb{X}_m,\mathbb{Y}_n\big) \right\}.$$

By assumption, we have $\mathbb{P}_{p\times q\times r}(\mathcal{B}) \geq 1-\frac{\beta}{2}$, and we want to show $\mathbb{P}_{p\times q\times r}(\mathcal{A}) \leq \beta$. Note that

$$\begin{aligned}
\mathbb{P}_{p\times q\times r}(\mathcal{A}\,|\,\mathcal{B}) &= \mathbb{P}_{p\times q\times r}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) \leq \widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)\ \Big|\ \mathcal{B}\Big) \\
&\leq \mathbb{P}_{p\times q}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) \leq \mathrm{MMD}^2_\lambda(p,q) - \sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big)}\right) \\
&= \mathbb{P}_{p\times q}\left(\mathrm{MMD}^2_\lambda(p,q) - \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) \geq \sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big)}\right) \\
&\leq \mathbb{P}_{p\times q}\left(\Big|\mathrm{MMD}^2_\lambda(p,q) - \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big| \geq \sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big)}\right) \\
&\leq \frac{\beta}{2}
\end{aligned}$$

by Chebyshev's inequality (Chebyshev, 1899) as $\mathbb{E}_{p\times q}\Big[\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big] = \mathrm{MMD}^2_\lambda(p,q)$. We then have

$$\begin{aligned}
\mathbb{P}_{p\times q\times r}(\mathcal{A}) &= \mathbb{P}_{p\times q\times r}(\mathcal{A}\,|\,\mathcal{B})\mathbb{P}_{p\times q\times r}(\mathcal{B}) + \mathbb{P}_{p\times q\times r}(\mathcal{A}\,|\,\mathcal{B}^c)\mathbb{P}_{p\times q\times r}(\mathcal{B}^c) \\
&\leq \frac{\beta}{2}\cdot 1 + 1\cdot\frac{\beta}{2} \\
&= \beta.
\end{aligned}$$

E.3 Proof of Proposition 3

We prove this result separately for our two MMD estimators $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ and $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$ defined in Equations (3) and (6), respectively.

Variance bound for MMD estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ defined in Equation (3).

In this case, we use the fact that $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ can be written as a two-sample U-statistic as in Equation (4). As noted by Kim et al. (2022, Appendix E, Part 1) one can derive from the explicit variance formula of the two-sample U-statistic (Lee, 1990, Equation 2 p.38) that there exists some positive constant c_0 such that

$$\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}(\mathbb{X}_m,\mathbb{Y}_n)\Big) \leq c_0\left(\frac{\sigma^2_{\lambda,1,0}}{m} + \frac{\sigma^2_{\lambda,0,1}}{n} + \left(\frac{1}{m}+\frac{1}{n}\right)^2\sigma^2_{\lambda,2,2}\right)$$

for

$$\begin{aligned}
\sigma^2_{\lambda,1,0} &\coloneqq \mathrm{var}_X\big(\mathbb{E}_{X',Y,Y'}\big[h_\lambda(X,X',Y,Y')\big]\big), \\
\sigma^2_{\lambda,0,1} &\coloneqq \mathrm{var}_Y\big(\mathbb{E}_{X,X',Y'}\big[h_\lambda(X,X',Y,Y')\big]\big), \\
\sigma^2_{\lambda,2,2} &\coloneqq \mathrm{var}_{X,X',Y,Y'}\big(h_\lambda(X,X',Y,Y')\big),
\end{aligned}$$

where $X, X' \overset{\text{iid}}{\sim} p$ and $Y, Y' \overset{\text{iid}}{\sim} q$ are all independent of each other. Making use of Equation (20), we deduce that there exists a positive constant $c_0^\dagger$ such that

$$\operatorname{var}_{p\times q}\left(\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}(\mathbb{X}_m, \mathbb{Y}_n)\right) \leq c_0^\dagger \left(\frac{\sigma^2_{\lambda,1,0} + \sigma^2_{\lambda,0,1}}{m+n} + \frac{\sigma^2_{\lambda,2,2}}{(m+n)^2}\right).$$

Recall that $\varphi_\lambda(u) \coloneqq \prod_{i=1}^d \frac{1}{\lambda_i} K_i\left(\frac{u_i}{\lambda_i}\right)$ for $u \in \mathbb{R}^d$ and that $\psi \coloneqq p - q$. Letting $G_\lambda = \psi * \varphi_\lambda$, we then have for all $u \in \mathbb{R}^d$

$$\begin{aligned}
G_\lambda(u) &= (\psi * \varphi_\lambda)(u) \\
&= \int_{\mathbb{R}^d} \psi(u')\varphi_\lambda(u-u')\mathrm{d}u' \\
&= \int_{\mathbb{R}^d} \psi(u')k_\lambda(u,u')\mathrm{d}u' \\
&= \int_{\mathbb{R}^d} k_\lambda(u,u')\left(p(u') - q(u')\right)\mathrm{d}u' \\
&= \mathbb{E}_{X'}\left[k_\lambda(u,X')\right] - \mathbb{E}_{Y'}\left[k_\lambda(u,Y')\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}_{X',Y'}\left[h_\lambda(X,X',Y,Y')\right] &= \mathbb{E}_{X',Y'}\left[k_\lambda(X,X') + k_\lambda(Y,Y') - k_\lambda(X,Y') - k_\lambda(X',Y)\right] \\
&= \mathbb{E}_{X'}\left[k_\lambda(X,X')\right] - \mathbb{E}_{Y'}\left[k_\lambda(X,Y')\right] \\
&\quad - \left(\mathbb{E}_{X'}\left[k_\lambda(Y,X')\right] - \mathbb{E}_{Y'}\left[k_\lambda(Y,Y')\right]\right) \\
&= G_\lambda(X) - G_\lambda(Y).
\end{aligned}$$

Hence, we get

$$\begin{aligned}
\sigma^2_{\lambda,1,0} &\coloneqq \operatorname{var}_X\left(\mathbb{E}_{X',Y,Y'}\left[h_\lambda(X,X',Y,Y')\right]\right) \\
&= \operatorname{var}_X(\mathbb{E}_Y[G_\lambda(X) - G_\lambda(Y)]) \\
&= \operatorname{var}_X(G_\lambda(X) - \mathbb{E}_Y[G_\lambda(Y)]) \\
&= \operatorname{var}_X(G_\lambda(X)) \\
&\leq \mathbb{E}_X\left[G_\lambda(X)^2\right] \\
&= \int_{\mathbb{R}^d} G_\lambda(x)^2 p(x)\mathrm{d}x \\
&\leq \|p\|_\infty \int_{\mathbb{R}^d} G_\lambda(x)^2 \mathrm{d}x \\
&\leq M\|G_\lambda\|_2^2 \\
&= M\|\psi * \varphi_\lambda\|_2^2
\end{aligned}$$

and, similarly, we get

$$\sigma^2_{\lambda,0,1} \coloneqq \operatorname{var}_Y\left(\mathbb{E}_{X,X',Y'}\left[h_\lambda(X,X',Y,Y')\right]\right) \leq M\|\psi * \varphi_\lambda\|_2^2.$$

For the third term, we have

$$\begin{aligned}\sigma^2_{\lambda,2,2} &\coloneqq \mathrm{var}_{X,X',Y,Y'}\big(h_\lambda(X,X',Y,Y')\big)\\ &= \mathrm{var}_{X,X',Y,Y'}\big(k_\lambda(X,X') + k_\lambda(Y,Y') - k_\lambda(X,Y') - k_\lambda(X',Y)\big)\\ &\leq \mathbb{E}_{X,X',Y,Y'}\Big[\big(k_\lambda(X,X') + k_\lambda(Y,Y') - k_\lambda(X,Y') - k_\lambda(X',Y)\big)^2\Big]\\ &\leq 4\big(\mathbb{E}_{X,X'}\big[k_\lambda(X,X')^2\big] + \mathbb{E}_{Y,Y'}\big[k_\lambda(Y,Y')^2\big] + 2\mathbb{E}_{X,Y}\big[k_\lambda(X,Y)^2\big]\big)\,.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}_{X,Y}\big[k_\lambda(X,Y)^2\big] &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_\lambda(x,y)^2 p(x)q(y)\,\mathrm{d}x\mathrm{d}y\\ &\leq \|p\|_\infty \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} k_\lambda(x,y)^2\mathrm{d}x\right) q(y)\,\mathrm{d}y\\ &= \|p\|_\infty \frac{\kappa_2}{\lambda_1\cdots\lambda_d}\int_{\mathbb{R}^d} q(y)\,\mathrm{d}y\\ &\leq \frac{M\kappa_2}{\lambda_1\cdots\lambda_d}\end{aligned}$$

where κ_2 depends on d and is defined in Equation (21). Similarly, we have

$$\mathbb{E}_{X,X'}\big[k_\lambda(X,X')^2\big] \leq \frac{M\kappa_2}{\lambda_1\cdots\lambda_d} \qquad \text{and} \qquad \mathbb{E}_{Y,Y'}\big[k_\lambda(Y,Y')^2\big] \leq \frac{M\kappa_2}{\lambda_1\cdots\lambda_d}. \tag{22}$$

We deduce that

$$\sigma^2_{\lambda,2,2} \coloneqq \mathrm{var}_{X,X',Y,Y'}\big(h_\lambda(X,X',Y,Y')\big) \leq \frac{16M\kappa_2}{\lambda_1\cdots\lambda_d}.$$

Letting $C_1(M,d) \coloneqq \max\left\{2c_0^\dagger M, 16c_0^\dagger M\kappa_2\right\}$ and combining the results, we obtain

$$\begin{aligned}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{a}}(\mathbb{X}_m,\mathbb{Y}_n)\Big) &\leq c_0^\dagger\left(\frac{\sigma^2_{\lambda,1,0}+\sigma^2_{\lambda,0,1}}{m+n} + \frac{\sigma^2_{\lambda,2,2}}{(m+n)^2}\right)\\ &\leq C_1(M,d)\left(\frac{\|\psi*\varphi_\lambda\|_2^2}{m+n} + \frac{1}{(m+n)^2\lambda_1\cdots\lambda_d}\right).\end{aligned}$$

Variance bound for MMD estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}$ defined in Equation (6).
The MMD estimator

$$\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}(\mathbb{X}_n,\mathbb{Y}_n) \coloneqq \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)$$

is a one-sample U-statistic of order 2. Hence, we can apply the result of Albert et al. (2022, Lemma 10) to get

$$\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_{\lambda,\mathsf{b}}(\mathbb{X}_n,\mathbb{Y}_n)\Big) \leq \widetilde{c}_0\left(\frac{\sigma^2_{\lambda,1,1}}{n} + \frac{\sigma^2_{\lambda,2,2}}{n^2}\right)$$

for some positive constant $\widetilde{c}_0$, where

$$\sigma^2_{\lambda,1,1} \coloneqq \mathrm{var}_{X,Y}\big(\mathbb{E}_{X',Y}\big[h_\lambda(X,X',Y,Y')\big]\big)$$

and

$$\sigma^2_{\lambda,2,2} \coloneqq \mathrm{var}_{X,X',Y,Y'}\big(h_\lambda(X,X',Y,Y')\big) \leq \frac{16M\kappa_2}{\lambda_1\cdots\lambda_d}$$

as shown earlier. Using the above results, we get

$$\begin{aligned}
\sigma^2_{\lambda,1,1} &\coloneqq \mathrm{var}_{X,Y}\big(\mathbb{E}_{X',Y}\big[h_\lambda(X,X',Y,Y')\big]\big)\\
&= \mathrm{var}_{X,Y}(G_\lambda(X) - G_\lambda(Y))\\
&\leq \mathbb{E}_{X,Y}\Big[\big(G_\lambda(X) - G_\lambda(Y)\big)^2\Big]\\
&\leq 2\big(\mathbb{E}_X\big[G_\lambda(X)^2\big] + \mathbb{E}_Y\big[G_\lambda(Y)^2\big]\big)\\
&\leq 4M\|\psi * \varphi_\lambda\|_2^2.
\end{aligned}$$

Letting $\widetilde{C}_1(M,d) \coloneqq 4\max\{4\widetilde{c}_0 M, 16\widetilde{c}_0 M\kappa_2\}$, we deduce that

$$\begin{aligned}
\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}(\mathbb{X}_n,\mathbb{Y}_n)\Big) &\leq \widetilde{c}_0\left(\frac{\sigma^2_{\lambda,1,1}}{n} + \frac{\sigma^2_{\lambda,2,2}}{n^2}\right)\\
&\leq \frac{1}{4}\widetilde{C}_1(M,d)\left(\frac{\|\psi * \varphi_\lambda\|_2^2}{n} + \frac{1}{n^2\lambda_1\cdots\lambda_d}\right)\\
&\leq \widetilde{C}_1(M,d)\left(\frac{\|\psi * \varphi_\lambda\|_2^2}{2n} + \frac{1}{(2n)^2\lambda_1\cdots\lambda_d}\right).
\end{aligned}$$

E.4 Proof of Proposition 4

Recall that $(\widehat{M}^b_\lambda)_{1\leq b\leq B}$ is defined in Sections 3.2.1 and 3.2.2 for the estimators $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ and $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$, respectively, and that $\widehat{M}^{B+1}_\lambda \coloneqq \widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)$ for both estimators. Let us recall that the $(1-\alpha)$-quantile function of a random variable X with cumulative distribution function F_X is given by

$$q_{1-\alpha} = \inf\{x\in\mathbb{R} : 1-\alpha \leq F_X(x)\}.$$

We denote by F_B and F_{B+1} the empirical cumulative distribution functions of $(\widehat{M}^b_\lambda)_{1\leq b\leq B}$ and $(\widehat{M}^b_\lambda)_{1\leq b\leq B+1}$, respectively.

For the case of the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$, we denote by F_∞ the cumulative distribution function of the conditional distribution of $\widehat{M}^\sigma_\lambda$ (defined in Equation (10)) given $\mathbb{X}_m$ and $\mathbb{Y}_n$, where the randomness comes from the uniform choice of permutation σ among all possible permutations of $\{1,\dots,m+n\}$. For the case of the estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$, we similarly denote by F_∞ the cumulative distribution function of the conditional distribution of $\widehat{M}^\epsilon_\lambda$ (defined in Equation (11)) given $\mathbb{X}_m$ and $\mathbb{Y}_n$, where the randomness comes from the n i.i.d. Rademacher variables $\epsilon \coloneqq (\epsilon_1,\dots,\epsilon_n)$ with values in $\{-1,1\}^n$.

Based on the above definitions, we can write

$$\widehat{q}^{\lambda,\infty}_{1-\alpha}(\mathbb{X}_m, \mathbb{Y}_n) = \inf\{u \in \mathbb{R} : 1-\alpha \leq F_\infty(u)\}$$

and

$$\begin{aligned}
\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B \big| \mathbb{X}_m, \mathbb{Y}_n\big) &= \inf\{u \in \mathbb{R} : 1-\alpha \leq F_{B+1}(u)\} \\
&= \inf\left\{u \in \mathbb{R} : 1-\alpha \leq \frac{1}{B+1}\sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}^b_\lambda \leq u\Big)\right\} \\
&= \inf\left\{u \in \mathbb{R} : (B+1)(1-\alpha) \leq \sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}^b_\lambda \leq u\Big)\right\} \\
&= \widehat{M}^{\bullet\lceil (B+1)(1-\alpha)\rceil}_\lambda
\end{aligned}$$

where $\widehat{M}^{\bullet 1}_\lambda \leq \cdots \leq \widehat{M}^{\bullet B+1}_\lambda$ denote the ordered simulated test statistics $(\widehat{M}^b_\lambda)_{1\leq b\leq B+1}$.

Now, for any given $\delta > 0$, define the event

$$\mathcal{A} \coloneqq \left\{\sup_{u\in\mathbb{R}} |F_B(u) - F_\infty(u)| \leq \sqrt{\frac{1}{2B}\log\left(\frac{4}{\delta}\right)}\right\}.$$

As noted by Kim et al. (2022, Remark 2.1), Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956), more precisely the version with the tight constant which is due to Massart (1990), then guarantees that $\mathbb{P}_r(\mathcal{A}|\mathbb{X}_m, \mathbb{Y}_n) \geq 1 - \frac{\delta}{2}$ for any $\mathbb{X}_m$ and $\mathbb{Y}_n$, so we deduce that $\mathbb{P}_{p\times q\times r}(\mathcal{A}) \geq 1 - \frac{\delta}{2}$. We now assume that the event $\mathcal{A}$ holds, so the bound we derive holds with probability $1 - \frac{\delta}{2}$. Notice that we cannot directly apply the Dvoretzky–Kiefer–Wolfowitz inequality to F_{B+1} since it is not based on i.i.d. samples. Nevertheless, under the event $\mathcal{A}$, we have

$$\begin{aligned}
\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B \big| \mathbb{X}_m, \mathbb{Y}_n\big) &= \inf\{u \in \mathbb{R} : 1-\alpha \leq F_{B+1}(u)\} \\
&= \inf\left\{u \in \mathbb{R} : 1-\alpha \leq \frac{1}{B+1}\sum_{b=1}^{B+1} \mathbb{1}\Big(\widehat{M}^b_\lambda \leq u\Big)\right\} \\
&\leq \inf\left\{u \in \mathbb{R} : 1-\alpha \leq \frac{1}{B+1}\sum_{b=1}^{B} \mathbb{1}\Big(\widehat{M}^b_\lambda \leq u\Big)\right\} \\
&= \inf\left\{u \in \mathbb{R} : (1-\alpha)\frac{B+1}{B} \leq F_B(u)\right\} \\
&\leq \inf\left\{u \in \mathbb{R} : (1-\alpha)\frac{B+1}{B} \leq F_\infty(u) - \sqrt{\frac{1}{2B}\log\left(\frac{4}{\delta}\right)}\right\} \\
&= \inf\left\{u \in \mathbb{R} : \underbrace{(1-\alpha)\frac{B+1}{B} + \sqrt{\frac{1}{2B}\log\left(\frac{4}{\delta}\right)}}_{\coloneqq 1-\alpha^*} \leq F_\infty(u)\right\} \\
&= \widehat{q}^{\lambda,\infty}_{1-\alpha^*}(\mathbb{X}_m, \mathbb{Y}_n).
\end{aligned}$$

Now, we take B large enough (only depending on α and δ) such that

$$(1-\alpha)\frac{B+1}{B} + \sqrt{\frac{1}{2B}\log\left(\frac{4}{\delta}\right)} \leq 1 - \frac{\alpha}{2}$$

so that $\widehat{q}^{\lambda,\infty}_{1-\alpha^*}(\mathbb{X}_m, \mathbb{Y}_n) \leq \widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_m, \mathbb{Y}_n)$ under the event $\mathcal{A}$. By reducing this problem to a quadratic equation with respect to B, we find

$$B \geq \frac{2}{\alpha^2}\left(\frac{1}{2}\ln\left(\frac{4}{\delta}\right) + \alpha - \alpha^2 + \sqrt{\left(\frac{1}{2}\ln\left(\frac{4}{\delta}\right) + \alpha - \alpha^2\right)^2 - \alpha^2(1-\alpha)^2}\right).$$

In particular, by upper bounding $-\alpha^2(1-\alpha)^2$ by 0, we find that the above inequality holds as soon as

$$B \geq \frac{4}{\alpha^2}\left(\frac{1}{2}\ln\left(\frac{4}{\delta}\right) + \alpha(1-\alpha)\right).$$

Note that this condition is in particular trivially satisfied if

$$B \geq \frac{3}{\alpha^2}\left(\ln\left(\frac{4}{\delta}\right) + \alpha(1-\alpha)\right).$$

With this choice of B, we have

$$\mathbb{P}_{p\times q\times r}\left(\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B\big|\mathbb{X}_m, \mathbb{Y}_n\big) \leq \widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_m, \mathbb{Y}_n)\right) \geq 1 - \frac{\delta}{2}.$$

We now upper bound $\widehat{q}^{\lambda,\infty}_{1-\alpha}(\mathbb{X}_m, \mathbb{Y}_n)$ for the two estimators $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ and $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{b}}$ separately. We then use this to prove the required upper bound on $\widehat{q}^{\lambda,B}_{1-\alpha}\big(\mathbb{Z}_B\big|\mathbb{X}_m, \mathbb{Y}_n\big)$.

Quantile bound for MMD estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathtt{a}}$ defined in Equation (3).

In this case, we base our reasoning on the work of Kim et al. (2022, proof of Lemma C.1). Recall from Equation (7) that we assume that $m \leq n$ and $n \leq Cm$ for some positive constant C. We use the notation presented in Section 3.2.1 that $U_i \coloneqq X_i$ and $U_{m+j} \coloneqq Y_j$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. By the result of Kim et al. (2022, Equation 59), there exists some $c_1 > 0$ such that

$$\widehat{q}^{\lambda,\infty}_{1-\alpha}(\mathbb{X}_m, \mathbb{Y}_n) \leq c_1\sqrt{\frac{1}{m^2(m-1)^2}\sum_{1\leq i\neq j\leq m+n} k_\lambda(U_i, U_j)^2}\ln\left(\frac{1}{\alpha}\right)$$

almost surely. As shown in Equation (22), for the constant $\kappa_2(d)$ defined in Equation (21), we have

$$\max\left\{\mathbb{E}_{X,X'}\left[k_\lambda(X,X')^2\right], \mathbb{E}_{X,Y}\left[k_\lambda(X,Y)^2\right], \mathbb{E}_{Y,Y'}\left[k_\lambda(Y,Y')^2\right]\right\} \leq \frac{M\kappa_2}{\lambda_1\cdots\lambda_d} \tag{23}$$

where $X, X' \overset{\text{iid}}{\sim} p$ and $Y, Y' \overset{\text{iid}}{\sim} q$ are all independent of each other. We deduce that

$$\begin{aligned}
\mathbb{E}_{p\times q}\left[\frac{1}{m^2(m-1)^2}\sum_{1\leq i\neq j\leq m+n} k_\lambda(U_i,U_j)^2\right] &\leq \frac{(m+n)(m+n-1)}{m^2(m-1)^2}\frac{M\kappa_2}{\lambda_1\cdots\lambda_d} \\
&\leq \frac{4M\kappa_2}{\lambda_1\cdots\lambda_d}\frac{(m+n)^2}{m^4} \\
&= \frac{64M\kappa_2}{\lambda_1\cdots\lambda_d}\frac{(m+n)^2}{(2m)^4} \\
&\leq \frac{64M\kappa_2}{\lambda_1\cdots\lambda_d}\frac{(m+n)^2}{(m+C^{-1}n)^4} \\
&\leq \frac{64M\kappa_2C^4}{\lambda_1\cdots\lambda_d}\frac{1}{(m+n)^2}
\end{aligned}$$

where we use the fact that $n \leq Cm$. Using Markov's inequality, we get that, for any $\delta \in (0,1)$, we have

$$\begin{aligned}
&1-\frac{\delta}{2} \\
\leq\ & \mathbb{P}_{p\times q}\left(\frac{1}{m^2(m-1)^2}\sum_{1\leq i\neq j\leq m+n} k_\lambda(U_i,U_j)^2 \leq \frac{2}{\delta}\mathbb{E}_{p\times q}\left[\frac{1}{m^2(m-1)^2}\sum_{1\leq i\neq j\leq m+n} k_\lambda(U_i,U_j)^2\right]\right) \\
\leq\ & \mathbb{P}_{p\times q}\left(\frac{1}{m^2(m-1)^2}\sum_{1\leq i\neq j\leq m+n} k_\lambda(U_i,U_j)^2 \leq \frac{2}{\delta}\frac{64M\kappa_2C^4}{(m+n)^2\lambda_1\cdots\lambda_d}\right) \\
\leq\ & \mathbb{P}_{p\times q}\left(\widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_m,\mathbb{Y}_n) \leq \frac{1}{\sqrt{\delta}}8c_1C^2\sqrt{2M\kappa_2}\frac{\ln\left(\frac{2}{\alpha}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right) \\
\leq\ & \mathbb{P}_{p\times q}\left(\widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_m,\mathbb{Y}_n) \leq \frac{1}{\sqrt{\delta}}16c_1C^2\sqrt{2M\kappa_2}\frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)
\end{aligned}$$

as $\ln\left(\frac{2}{\alpha}\right) \leq 2\ln\left(\frac{1}{\alpha}\right)$ since $\alpha \in (0, 0.5)$. We now let $C_2(M,d) \coloneqq 16c_1C^2\sqrt{2M\kappa_2}$. Then, for all $B \in \mathbb{N}$ such that $B \geq \frac{3}{\alpha^2}\left(\ln\left(\frac{4}{\delta}\right)+\alpha(1-\alpha)\right)$, we have

$$\begin{aligned}
&\mathbb{P}_{p\times q\times r}\left(\widehat{q}^{\lambda,B}_{1-\alpha}\left(\mathbb{Z}_B\middle|\mathbb{X}_m,\mathbb{Y}_n\right) \leq \frac{1}{\sqrt{\delta}}C_2(M,d)\frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right) \\
\geq\ & \mathbb{P}_{p\times q\times r}\Bigg(\left\{\widehat{q}^{\lambda,B}_{1-\alpha}\left(\mathbb{Z}_B\middle|\mathbb{X}_m,\mathbb{Y}_n\right) \leq \widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_m,\mathbb{Y}_n)\right\} \\
&\qquad \cap \left\{\widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_m,\mathbb{Y}_n) \leq \frac{1}{\sqrt{\delta}}C_2(M,d)\frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right\}\Bigg) \\
\geq\ & 1-\frac{\delta}{2}-\frac{\delta}{2} \\
=\ & 1-\delta
\end{aligned}$$

where we use the standard fact that for events $\mathcal{B}$ and $\mathcal{C}$ satisfying $\mathbb{P}(\mathcal{B}) \geq 1-\delta_1$ and $\mathbb{P}(\mathcal{C}) \geq 1-\delta_2$, we have $\mathbb{P}(\mathcal{B}\cap\mathcal{C}) = 1-\mathbb{P}(\mathcal{B}^c\cup\mathcal{C}^c) \geq 1-\mathbb{P}(\mathcal{B}^c)-\mathbb{P}(\mathcal{C}^c) \geq 1-\delta_1-\delta_2$.

Quantile bound for MMD estimator $\widehat{\mathrm{MMD}}^2_{\lambda,\mathfrak{b}}$ defined in Equation (6).

For this case, in order to upper bound $\widehat{q}^{\lambda,\infty}_{1-\alpha}(\mathbb{X}_n,\mathbb{Y}_n)$, we can use the result of de la Peña and Giné (1999, Corollary 3.2.6) and Markov's inequality as done by Fromont et al. (2012, Appendix D). We obtain that there exists a positive constant $\widetilde{c}_1$ such that

$$\mathbb{P}_r\left(\left|\sum_{1\leq i\neq j\leq n}\epsilon_i\epsilon_j h_\lambda(X_i,X_j,Y_i,Y_j)\right| \geq \widetilde{c}_1\sqrt{\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)^2}\ln\left(\frac{2}{\alpha}\right)\,\middle|\,\mathbb{X}_n,\mathbb{Y}_n\right)\leq\alpha$$

where $\epsilon_1,\ldots,\epsilon_n$ are i.i.d. Rademacher variables, and so $\sum_{1\leq i\neq j\leq n}\epsilon_i\epsilon_j h_\lambda(X_i,X_j,Y_i,Y_j)$ is a Rademacher chaos. We deduce that

$$\widehat{q}^{\lambda,\infty}_{1-\alpha}(\mathbb{X}_n,\mathbb{Y}_n) \leq \widetilde{c}_1\sqrt{\frac{1}{n^2(n-1)^2}\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)^2}\ln\left(\frac{2}{\alpha}\right)$$

almost surely. Using Equation (23), we have

$$\begin{aligned}
&\mathbb{E}_{p\times q}\left[\frac{1}{n^2(n-1)^2}\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)^2\right]\\
=\;&\frac{1}{n^2(n-1)^2}\mathbb{E}_{p\times q}\left[\sum_{1\leq i\neq j\leq n}\big(k(X_i,X_j)+k(Y_i,Y_j)-k(X_i,Y_j)-k(X_j,Y_i)\big)^2\right]\\
\leq\;&\frac{4}{n(n-1)}\Big(\mathbb{E}_{X,X'}\big[k_\lambda(X,X')^2\big]+\mathbb{E}_{Y,Y'}\big[k_\lambda(Y,Y')^2\big]+2\mathbb{E}_{X,Y}\big[k_\lambda(X,Y)^2\big]\Big)\\
\leq\;&\frac{32M\kappa_2}{n^2\lambda_1\cdots\lambda_d}
\end{aligned}$$

where $X,X' \overset{\text{iid}}{\sim} p$ and $Y,Y' \overset{\text{iid}}{\sim} q$ are all independent of each other, and where κ_2 is the constant defined in Equation (21) depending on d. Similarly to the previous case, we can then use Markov's inequality to get that, for any $\delta\in(0,1)$, we have

$$\begin{aligned}
&1-\frac{\delta}{2}\\
&\leq \mathbb{P}_{p\times q}\left(\frac{1}{n^2(n-1)^2}\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)^2 \leq \frac{2}{\delta}\mathbb{E}_{p\times q}\left[\frac{1}{n^2(n-1)^2}\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)^2\right]\right)\\
&\leq \mathbb{P}_{p\times q}\left(\frac{1}{n^2(n-1)^2}\sum_{1\leq i\neq j\leq n} h_\lambda(X_i,X_j,Y_i,Y_j)^2 \leq \frac{2}{\delta}\frac{32M\kappa_2}{n^2\lambda_1\cdots\lambda_d}\right)\\
&\leq \mathbb{P}_{p\times q}\left(\widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_n,\mathbb{Y}_n) \leq \frac{1}{\sqrt{\delta}}8\widetilde{c}_1\sqrt{M\kappa_2}\frac{\ln\left(\frac{4}{\alpha}\right)}{n\sqrt{\lambda_1\cdots\lambda_d}}\right)\\
&\leq \mathbb{P}_{p\times q}\left(\widehat{q}^{\lambda,\infty}_{1-\alpha/2}(\mathbb{X}_n,\mathbb{Y}_n) \leq \frac{1}{\sqrt{\delta}}48\widetilde{c}_1\sqrt{M\kappa_2}\frac{\ln\left(\frac{1}{\alpha}\right)}{2n\sqrt{\lambda_1\cdots\lambda_d}}\right)
\end{aligned}$$

as $\ln\left(\frac{4}{\alpha}\right) \leq 3\ln\left(\frac{1}{\alpha}\right)$ since $\alpha \in (0, 0.5)$. Letting $\widetilde{C}_2(M,d) \coloneqq 48\widetilde{c}_1\sqrt{M\kappa_2}$ and applying the same reasoning as earlier, we get

$$\mathbb{P}_{p\times q\times r}\left(\widehat{q}^{\lambda,B}_{1-\alpha}\left(\mathbb{Z}_B\middle|\mathbb{X}_m,\mathbb{Y}_n\right) \leq \frac{1}{\sqrt{\delta}}\widetilde{C}_2(M,d)\frac{\ln\left(\frac{1}{\alpha}\right)}{2n\sqrt{\lambda_1\cdots\lambda_d}}\right) \leq 1-\delta$$

for all $B \in \mathbb{N}$ satisfying $B \geq \frac{3}{\alpha^2}\left(\ln\left(\frac{4}{\delta}\right) + \alpha(1-\alpha)\right)$.

E.5 Proof of Theorem 5

First, as shown by Gretton et al. (2012a, Lemma 6), the Maximum Mean Discrepancy can be written as

$$\begin{aligned}
\mathrm{MMD}^2_\lambda(p,q) &= \mathbb{E}_{X,X'}\left[k_\lambda(X,X')\right] - 2\,\mathbb{E}_{X,Y}\left[k_\lambda(X,Y)\right] + \mathbb{E}_{Y,Y'}\left[k_\lambda(Y,Y')\right] \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_\lambda(x,x')p(x)p(x')\,\mathrm{d}x\mathrm{d}x' \\
&\quad - 2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_\lambda(x,y)p(x)q(y)\,\mathrm{d}x\mathrm{d}y \\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_\lambda(y,y')q(y)q(y')\,\mathrm{d}y\mathrm{d}y' \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_\lambda(u,u')\big(p(u)-q(u)\big)\big(p(u')-q(u')\big)\,\mathrm{d}u\mathrm{d}u'
\end{aligned}$$

for $X, X' \overset{\text{iid}}{\sim} p$ and $Y, Y' \overset{\text{iid}}{\sim} q$ all independent of each other. Using the function φ_λ defined in Equation (9) and $\psi \coloneqq p - q$, we obtain

$$\begin{aligned}
\mathrm{MMD}^2_\lambda(p,q) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \varphi_\lambda(u-u')\psi(u)\psi(u')\,\mathrm{d}u\mathrm{d}u' \\
&= \int_{\mathbb{R}^d}\psi(u)\int_{\mathbb{R}^d}\psi(u')\varphi_\lambda(u-u')\,\mathrm{d}u'\mathrm{d}u \\
&= \int_{\mathbb{R}^d}\psi(u)\big(\psi * \varphi_\lambda\big)(u)\,\mathrm{d}u \\
&= \langle\psi, \psi*\varphi_\lambda\rangle_2 \\
&= \frac{1}{2}\Big(\|\psi\|_2^2 + \|\psi*\varphi_\lambda\|_2^2 - \|\psi-\psi*\varphi_\lambda\|_2^2\Big)
\end{aligned}$$

where the last equality is obtained by expanding $\|\psi - \psi*\varphi_\lambda\|_2^2$. By Lemma 2, a sufficient condition to ensure that $\mathbb{P}_{p\times q\times r}\Big(\Delta^{\lambda,B}_\alpha(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_B)=0\Big) \leq \beta$ is

$$\mathbb{P}_{p\times q\times r}\left(\mathrm{MMD}^2_\lambda(p,q) \geq \sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\Big)} + \widehat{q}^{\lambda,B}_{1-\alpha}\left(\mathbb{Z}_B\middle|\mathbb{X}_m,\mathbb{Y}_n\right)\right) \geq 1-\frac{\beta}{2}$$

and an equivalent sufficient condition is

$$\mathbb{P}_{p\times q\times r}\left(\|\psi\|_2^2 \geq \|\psi-\psi*\varphi_\lambda\|_2^2 - \|\psi*\varphi_\lambda\|_2^2 + 2\sqrt{\frac{2}{\beta}\mathrm{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}^2_\lambda\Big)} + 2\widehat{q}^{\lambda,B}_{1-\alpha}\right) \geq 1-\frac{\beta}{2}.$$

By Proposition 3, we have

$$\begin{aligned}
\operatorname{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)\Big) &\le C_1(M,d)\left(\frac{\|\psi*\varphi_\lambda\|_2^2}{m+n}+\frac{1}{(m+n)^2\,\lambda_1\cdots\lambda_d}\right)\\
2\sqrt{\frac{2}{\beta}\operatorname{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)\Big)} &\le 2\sqrt{\frac{2C_1}{\beta}\frac{\|\psi*\varphi_\lambda\|_2^2}{m+n}+\frac{2C_1}{\beta(m+n)^2\,\lambda_1\cdots\lambda_d}}\\
&\le 2\sqrt{\|\psi*\varphi_\lambda\|_2^2\frac{2C_1}{\beta(m+n)}}+\frac{2\sqrt{2C_1}}{\sqrt{\beta}(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\\
&\le \|\psi*\varphi_\lambda\|_2^2+\frac{2C_1}{\beta(m+n)}+\frac{2\sqrt{2C_1}}{\sqrt{\beta}(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\\
&\le \|\psi*\varphi_\lambda\|_2^2+\frac{2C_1+2\sqrt{2C_1}}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\ln\left(\frac{1}{\alpha}\right)\\
&\le \|\psi*\varphi_\lambda\|_2^2+\frac{6C_1}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\ln\left(\frac{1}{\alpha}\right)\\
\|\psi*\varphi_\lambda\|_2^2-2\sqrt{\frac{2}{\beta}\operatorname{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}_\lambda^2\Big)} &\ge -6C_1\frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}
\end{aligned}$$

where for the third inequality we used the fact that $\sqrt{x+y}\le\sqrt{x}+\sqrt{y}$ for all $x,y>0$, for the fourth inequality we used the fact that $2\sqrt{xy}\le x+y$ for all $x,y>0$, and for the fifth inequality we use the fact that $\lambda_1\cdots\lambda_d\le 1$, $\beta\in(0,1)$ and $\ln\left(\frac{1}{\alpha}\right)>1$. A similar reasoning has been used by Fromont et al. (2013, Theorem 1) and Albert et al. (2022, Theorem 1).

Let $C_3(M,d)\coloneqq 6C_1(M,d)+2\sqrt{2}C_2(M,d)$ where C_1 and C_2 are the constants from Propositions 3 and 4, respectively. Assume that our condition holds, that is

$$\|\psi\|_2^2-\|\psi-\psi*\varphi_\lambda\|_2^2\ge(2\sqrt{2}C_2+6C_1)\frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}.$$

Omitting the variables for $\widehat{q}_{1-\alpha}^{\lambda,B}\big(\mathbb{Z}_B\big|\mathbb{X}_m,\mathbb{Y}_n\big)$ and for $\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m,\mathbb{Y}_n)$, we then get

$$\begin{aligned}
&\mathbb{P}_{p\times q\times r}\left(2\widehat{q}_{1-\alpha}^{\lambda,B}\le\|\psi\|_2^2-\|\psi-\psi*\varphi_\lambda\|_2^2+\|\psi*\varphi_\lambda\|_2^2-2\sqrt{\frac{2}{\beta}\operatorname{var}_{p\times q}\Big(\widehat{\mathrm{MMD}}_\lambda^2\Big)}\right)\\
&\ge\mathbb{P}_{p\times q\times r}\left(2\widehat{q}_{1-\alpha}^{\lambda,B}\le(6C_1+2\sqrt{2}C_2)\frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}-6C_1\frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\\
&=\mathbb{P}_{p\times q\times r}\left(\widehat{q}_{1-\alpha}^{\lambda,B}\le\sqrt{2}C_2\frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\\
&\ge\mathbb{P}_{p\times q\times r}\left(\widehat{q}_{1-\alpha}^{\lambda,B}\le C_2\sqrt{\frac{2}{\beta}}\frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\\
&\ge 1-\frac{\beta}{2}
\end{aligned}$$

where the third inequality holds because $\beta \in (0,1)$ and the last one holds by Proposition 4 since $B \geq \frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$. Lemma 2 then implies that

$$\mathbb{P}_{p\times q\times r}\Big(\Delta_\alpha^{\lambda,B}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_B) = 0\Big) \leq \beta.$$

E.6 Proof of Theorem 6

Theorem 5 gives us a condition on $\|\psi\|_2^2 - \|\psi - \psi * \varphi_\lambda\|_2^2$ to control the power of the test $\Delta_\alpha^{\lambda,B}$. We now want to upper bound $\|\psi - \psi * \varphi_\lambda\|_2^2$ in terms of the bandwidths when assuming that the difference of the densities lie in a Sobolev ball. We first prove that if $\psi \coloneqq p - q \in \mathcal{S}_d^s(R)$ for some $s > 0$ and $R > 0$, then there exists some $S \in (0,1)$ such that

$$\|\psi - \psi * \varphi_\lambda\|_2^2 - S^2\|\psi\|_2^2 \leq C_4'(d,s,R)\sum_{i=1}^d \lambda_i^{2s} \tag{24}$$

for some positive constant $C_4'(d,s,R)$.

For $j = 1,\dots,d$, since $K_j \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, it follows by the Riemann-Lebesgue Lemma that its Fourier transform $\widehat{K}_j$ is continuous. For $j = 1,\dots,d$, note that

$$\widehat{K}_j(0) = \int_{\mathbb{R}} K_j(x)e^{-ix0}\mathrm{d}x = \int_{\mathbb{R}} K_j(x)\mathrm{d}x = 1$$

and, since $K_j \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, also that

$$\prod_{j=1}^d \Big|\widehat{K}_j(\xi_j)\Big| \leq \prod_{j=1}^d \int_{\mathbb{R}} \Big|K_j(x)e^{-ix\xi_j}\Big|\mathrm{d}x = \prod_{j=1}^d \int_{\mathbb{R}} |K_j(x)|\mathrm{d}x =: \kappa_1 < \infty$$

as defined in Equation (21). We deduce that $\Big|1 - \prod_{i=1}^d \widehat{K}_i(\xi_i)\Big| \leq 1 + \kappa_1$ for all $\xi \in \mathbb{R}^d$. Let us define $g\colon \mathbb{R}^d \to \mathbb{R}$ by $g(\xi) = 1 - \prod_{i=1}^d \widehat{K}_i(\xi_i)$ for $\xi \in \mathbb{R}^d$. We have $g(0,\dots,0) = 0$, so by continuity of g, there exists some $t > 0$ such that

$$S \coloneqq \sup_{\|\xi\|_2 \leq t} |g(\xi)| < 1.$$

For any $s > 0$, we also define

$$T_s \coloneqq \sup_{\|\xi\|_2 > t} \frac{\Big|1 - \prod_{i=1}^d \widehat{K}_i(\xi_i)\Big|}{\|\xi\|_2^s} \leq \frac{1+\kappa_1}{t^s} < \infty.$$

Let $\Psi \coloneqq \psi - \psi * \varphi_\lambda$. As it is a scaled product of $K_1,\dots,K_d \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, we have $\varphi_\lambda \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Since we assume that $\psi \in \mathcal{S}_d^s(R)$, we have $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. For $p \in \{1,2\}$, since $\psi \in L^1(\mathbb{R}^d)$, we have $\|\psi * \varphi_\lambda\|_p \leq \|\psi\|_1 \|\varphi_\lambda\|_p < \infty$. Hence, we deduce that $\Psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. By Plancherel's Theorem, we then have

$$(2\pi)^d\|\Psi\|_2^2 = \|\widehat{\Psi}\|_2^2$$
$$(2\pi)^d\|\psi - \psi * \varphi_\lambda\|_2^2 = \Big\|(1 - \widehat{\varphi_\lambda})\widehat{\psi}\Big\|_2^2.$$

In general, for $a > 0$ the Fourier transform of a function $x \mapsto \frac{1}{a} f\left(\frac{x}{a}\right)$ is $\xi \mapsto \widehat{f}(a\xi)$. Since $\varphi_\lambda(u) \coloneqq \prod_{i=1}^d \frac{1}{\lambda_i} K_i\left(\frac{u_i}{\lambda_i}\right)$ for $u \in \mathbb{R}^d$, we deduce that $\widehat{\varphi_\lambda}(\xi) = \prod_{i=1}^d \widehat{K}_i(\lambda_i \xi_i)$ for $\xi \in \mathbb{R}^d$. Therefore, we have

$$
\begin{aligned}
&(2\pi)^d \|\psi - \psi * \varphi_\lambda\|_2^2 \\
&= \left\| (1 - \widehat{\varphi_\lambda})\widehat{\psi} \right\|_2^2 \\
&= \int_{\mathbb{R}^d} |1 - \widehat{\varphi_\lambda}(\xi)|^2 \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&= \int_{\mathbb{R}^d} \left| 1 - \prod_{i=1}^d \widehat{K}(\lambda_i \xi_i) \right|^2 \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&= \int_{\|\xi\|_2 \leq t} \left| 1 - \prod_{i=1}^d \widehat{K}(\lambda_i \xi_i) \right|^2 \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi + \int_{\|\xi\|_2 > t} \left| 1 - \prod_{i=1}^d \widehat{K}(\lambda_i \xi_i) \right|^2 \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&\leq S^2 \int_{\|\xi\|_2 \leq t} \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi + T_s^2 \int_{\|\xi\|_2 > t} \|(\lambda_1 \xi_1, \ldots, \lambda_d \xi_d)\|_2^{2s} \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&\leq S^2 \left\|\widehat{\psi}\right\|_2^2 + T_s^2 \int_{\mathbb{R}^d} \left(\sum_{i=1}^d \lambda_i^2 \xi_i^2 \right)^s \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&\leq S^2 (2\pi)^d \|\psi\|_2^2 + T_s^2 \int_{\mathbb{R}^d} \left(\sum_{i=1}^d \lambda_i^2 \right)^s \left(\sum_{i=1}^d \xi_i^2 \right)^s \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&= S^2 (2\pi)^d \|\psi\|_2^2 + T_s^2 \|\lambda\|_2^{2s} \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \\
&\leq S^2 (2\pi)^d \|\psi\|_2^2 + T_s^2 \|\lambda\|_2^{2s} (2\pi)^d R^2
\end{aligned}
$$

since $\psi \in \mathcal{S}_d^s(R)$, and where we have used Plancherel's Theorem for $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. We have proved that there exists some $S \in (0,1)$ such that

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

If $s \geq 1$, then $x \mapsto x^s$ is convex and, by Jensen's inequality (finite form), we have

$$
\|\lambda\|_2^{2s} = \left(\sum_{i=1}^d \lambda_i^2 \right)^s = d^s \left(\sum_{i=1}^d \frac{1}{d} \lambda_i^2 \right)^s \leq d^s \sum_{i=1}^d \frac{1}{d} \left(\lambda_i^2\right)^s = d^{s-1} \sum_{i=1}^d \lambda_i^{2s} \leq d^{1+s} \sum_{i=1}^d \lambda_i^{2s}.
$$

If $s < 1$, then $\gamma \coloneqq \frac{1}{s} > 1$ and so, it is a standard result that $\|\cdot\|_\gamma \leq \|\cdot\|_1$. We then have

$$
\|\lambda\|_2^{2s} = \left(\sum_{i=1}^d \lambda_i^2 \right)^s = \left(\sum_{i=1}^d \left(\lambda_i^{2s}\right)^\gamma \right)^{1/\gamma} = \left\|\lambda^{2s}\right\|_\gamma \leq \left\|\lambda^{2s}\right\|_1 = \sum_{i=1}^d \lambda_i^{2s} \leq d^{1+s} \sum_{i=1}^d \lambda_i^{2s}.
$$

Hence, for all $s > 0$, we have $\|\lambda\|_2^{2s} \leq d^{1+s} \sum_{i=1}^d \lambda_i^{2s}$. We conclude that

$$
\|\psi - \psi * \varphi_\lambda\|_2^2 \leq S^2 \|\psi\|_2^2 + T_s^2 R^2 d^{1+s} \sum_{i=1}^d \lambda_i^{2s}
$$

which proves the statement presented in Equation (24) with $C_4'(d,s,R) \coloneqq T_s^2 R^2 d^{1+s}$.

We now consider the constant $C_3(M,d)$ from Theorem 5. Suppose we have

$$\begin{aligned}
\|\psi\|_2^2 &\geq \frac{T_s^2 R^2 d^{1+s}}{1-S^2} \sum_{i=1}^d \lambda_i^{2s} + \frac{C_3}{(1-S^2)} \frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1 \cdots \lambda_d}} \\
(1-S^2)\|\psi\|_2^2 &\geq T_s^2 R^2 d^{1+s} \sum_{i=1}^d \lambda_i^{2s} + C_3 \frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1 \cdots \lambda_d}} \\
\|\psi\|_2^2 &\geq S^2 \|\psi\|_2^2 + T_s^2 R^2 d^{1+s} \sum_{i=1}^d \lambda_i^{2s} + C_3 \frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1 \cdots \lambda_d}} \\
\|\psi\|_2^2 &\geq \|\psi - \psi * \varphi_\lambda\|_2^2 + C_3 \frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1 \cdots \lambda_d}}
\end{aligned}$$

then, by Theorem 5, we can ensure that

$$\mathbb{P}_{p\times q\times r}\Big(\Delta_\alpha^{\lambda,B}(\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_B) = 0\Big) \leq \beta.$$

By definition of uniform separation rates, we deduce that

$$\begin{aligned}
\rho\Big(\Delta_\alpha^{\lambda,B}, \mathcal{S}_d^s(R), \beta, M\Big)^2 &\leq \frac{T_s^2 R^2 d^{1+s}}{1-S^2} \sum_{i=1}^d \lambda_i^{2s} + \frac{C_3}{(1-S^2)} \frac{\ln\left(\frac{1}{\alpha}\right)}{\beta(m+n)\sqrt{\lambda_1 \cdots \lambda_d}} \\
&\leq C_4(M,d,s,R,\beta)\left(\sum_{i=1}^d \lambda_i^{2s} + \frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1 \cdots \lambda_d}}\right)
\end{aligned}$$

for $C_4(M,d,s,R,\beta) \coloneqq \max\left\{\frac{T_s^2 R^2 d^{1+s}}{1-S^2}, \frac{C_3(M,d)}{\beta(1-S^2)}\right\}$.

E.7 Proof of Corollary 7

By Theorem 6, if $\lambda_1 \cdots \lambda_d \leq 1$, we have

$$\rho\Big(\Delta_\alpha^{\lambda,B}, \mathcal{S}_d^s(R), \beta, M\Big)^2 \leq C_4(M,d,s,R,\beta)\left(\sum_{i=1}^d \lambda_i^{2s} + \frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1 \cdots \lambda_d}}\right).$$

We want to express the bandwidths in terms of the sum of sample sizes $m+n$ raised to some negative power such that the terms $\sum_{i=1}^d \lambda_i^{2s}$ and $\frac{1}{(m+n)\sqrt{\lambda_1 \cdots \lambda_d}}$ have the same behaviour in $m+n$. With the choice of bandwidths $\lambda_i^* \coloneqq (m+n)^{-2/(4s+d)}$ for $i = 1, \ldots, d$, the term $\sum_{i=1}^d (\lambda_i^*)^{2s}$ has order $(m+n)^{-4s/(4s+d)}$ and the term $\frac{1}{(m+n)\sqrt{\lambda_1^* \cdots \lambda_d^*}}$ has order $(m+n)^{d/(4s+d)-1} = (m+n)^{-4s/(4s+d)}$. So, indeed, this choice of bandwidths leads to the same behaviour in $m+n$ for the two terms, which gives the smallest order of $m+n$ possible.

It is clear that $\lambda_1^* \cdots \lambda_d^* < 1$, we find that

$$\begin{aligned}
\rho\Big(\Delta_\alpha^{\lambda^*,B}, \mathcal{S}_d^s(R), \beta, M\Big)^2 &\leq C_4(M,d,s,R,\beta)\left(\sum_{i=1}^d (\lambda_i^*)^{2s} + \frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1^* \cdots \lambda_d^*}}\right) \\
&\leq C_4(M,d,s,R,\beta)\left((m+n)^{-4s/(4s+d)} + \ln\left(\frac{1}{\alpha}\right)(m+n)^{-4s/(4s+d)}\right) \\
&\leq C_4(M,d,s,R,\beta)\ln\left(\frac{1}{\alpha}\right)(m+n)^{-4s/(4s+d)} \\
&= C_5(M,d,s,R,\alpha,\beta)^2\,(m+n)^{-4s/(4s+d)}
\end{aligned}$$

for $C_5(M,d,s,R,\alpha,\beta) \coloneqq \sqrt{C_4(M,d,s,R,\beta)\ln\left(\frac{1}{\alpha}\right)}$. We deduce that

$$\rho\Big(\Delta_\alpha^{\lambda^*,B}, \mathcal{S}_d^s(R), \beta, M\Big) \leq C_5(M,d,s,R,\alpha,\beta)\,(m+n)^{-2s/(4s+d)}\,.$$

E.8 Proof of Proposition 8

By definition of $u_\alpha^{B_2}$, we have

$$\frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}_{\lambda,2}^b\Big(\mu^{(b,2)}\Big|\mathbb{X}_m,\mathbb{Y}_n\Big) - \widehat{q}^{\lambda,B_1}_{1-u_\alpha^{B_2}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)w_\lambda}\Big(\mathbb{Z}_{B_1}\Big|\mathbb{X}_m,\mathbb{Y}_n\Big)\right) > 0\right) \leq \alpha.$$

Taking the expectation on both sides, we get

$$\mathbb{P}_{p\times p\times r\times r}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) - \widehat{q}^{\lambda,B_1}_{1-u_\alpha^{B_2}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)w_\lambda}\Big(\mathbb{Z}_{B_1}\Big|\mathbb{X}_m,\mathbb{Y}_n\Big)\right) > 0\right) \leq \alpha$$

as under the null hypothesis $\mathcal{H}_0\colon p = q$, we have $\big(\widehat{M}_{\lambda,2}^b\big(\mu^{(b,2)}\big|\mathbb{X}_m,\mathbb{Y}_n\big)\big)_{1\leq b\leq B_2}$ distributed like $\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)$. Using the bisection search approximation which satisfies

$$\widehat{u}_\alpha^{B_{2:3}}\big(\mathbb{Z}_{B_2}\big|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\big) \leq u_\alpha^{B_2}\big(\mathbb{Z}_{B_2}\big|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\big),$$

we get

$$\begin{aligned}
&\mathbb{P}_{p\times p\times r\times r}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) - \widehat{q}^{\lambda,B_1}_{1-\widehat{u}_\alpha^{B_{2:3}}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)w_\lambda}\Big(\mathbb{Z}_{B_1}\Big|\mathbb{X}_m,\mathbb{Y}_n\Big)\right) > 0\right) \\
&\leq \mathbb{P}_{p\times p\times r\times r}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) - \widehat{q}^{\lambda,B_1}_{1-u_\alpha^{B_2}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)w_\lambda}\Big(\mathbb{Z}_{B_1}\Big|\mathbb{X}_m,\mathbb{Y}_n\Big)\right) > 0\right) \\
&\leq \alpha.
\end{aligned}$$

We deduce that

$$\mathbb{P}_{p\times p\times r\times r}\big(\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1},\mathbb{Z}_{B_2}) = 1\big) \leq \alpha.$$

E.9 Proof of Theorem 9

Consider some $u^* \in (0,1)$ to be determined later. For $b = 1, \ldots, B_2$, let

$$W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big) \coloneqq \mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2}\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n\Big) - \widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\right) > 0\right),$$

so that, following a similar argument to the one presented in Appendix E.8, we obtain that $\mathbb{E}_{p\times q\times r\times r}\big[W_b\big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\big)\big]$ is equal to

$$\mathbb{P}_{p\times p\times r}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\right) > 0\right).$$

Consider the events

$$\mathcal{A}' \coloneqq \left\{\frac{1}{B_2}\sum_{b=1}^{B_2} W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big) - \mathbb{E}_r\Big[W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big)\Big] \leq \sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}\right\}$$

and

$$\mathcal{A} \coloneqq \left\{\frac{1}{B_2}\sum_{b=1}^{B_2} W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big) - \mathbb{E}_{p\times q\times r\times r}\Big[W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big)\Big] \leq \sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}\right\}.$$

Using Hoeffding's inequality, we obtain that $\mathbb{P}_r\big(\mathcal{A}'\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\big) \geq 1 - \frac{\beta}{2}$ for any $\mathbb{X}_m$, $\mathbb{Y}_n$ and $\mathbb{Z}_{B_1}$, we deduce that $\mathbb{P}_{p\times q\times r\times r}(\mathcal{A}) \geq 1 - \frac{\beta}{2}$.

First, assuming that the event $\mathcal{A}$ holds, we show that $u^{B_2}_\alpha\big(\mathbb{Z}_{B_2}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\big) \geq \alpha$. Since we assume that the event $\mathcal{A}$ holds, the bounds we obtain hold with probability $1 - \frac{\beta}{2}$. We have

$$\begin{aligned}
&\frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2}\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n\Big) - \widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\right) > 0\right) \\
&= \frac{1}{B_2}\sum_{b=1}^{B_2} W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big) \\
&\leq \mathbb{E}_{p\times p\times r\times r}\Big[W_b\Big(\mu^{(b,2)}\big|\mathbb{X}_m, \mathbb{Y}_n, \mathbb{Z}_{B_1}\Big)\Big] + \sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)} \\
&= \mathbb{P}_{p\times p\times r}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\right) > 0\right) + \sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)} \\
&= \mathbb{P}_{p\times p\times r}\left(\bigcup_{\lambda\in\Lambda}\left\{\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m, \mathbb{Y}_n\big)\right\}\right) + \sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}.
\end{aligned}$$

Using Boole's inequality, we obtain

$$\begin{aligned}
&\frac{1}{B_2}\sum_{b=1}^{B_2}\mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2}\left(\mu^{(b,2)}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)-\widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)\right)>0\right)\\
&\leq \sum_{\lambda\in\Lambda}\mathbb{P}_{p\times p\times r}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)>\widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)\right)+\sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}\\
&\leq \sum_{\lambda\in\Lambda}u^*w_\lambda+\sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}\\
&\leq u^*+\sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}\\
&= \frac{3\alpha}{4}+\sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}
\end{aligned}$$

for $u^* := \frac{3\alpha}{4}$, where we have used Proposition 1 and the fact that $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$. Now, for $B_2 \geq \frac{8}{\alpha^2}\ln\left(\frac{2}{\beta}\right)$, we get

$$\frac{3\alpha}{4}+\sqrt{\frac{1}{2B_2}\ln\left(\frac{2}{\beta}\right)}\leq\alpha$$

and so, we obtain

$$\frac{1}{B_2}\sum_{b=1}^{B_2}\mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2}\left(\mu^{(b,2)}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)-\widehat{q}^{\lambda,B_1}_{1-u^*w_\lambda}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)\right)>0\right)\leq\alpha.$$

Recall that $u^{B_2}_\alpha\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)$ is defined as

$$\sup\left\{u\in\left(0,\min_{\lambda\in\Lambda}w_\lambda^{-1}\right):\frac{1}{B_2}\sum_{b=1}^{B_2}\mathbb{1}\left(\max_{\lambda\in\Lambda}\left(\widehat{M}^b_{\lambda,2}\left(\mu^{(b,2)}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)-\widehat{q}^{\lambda,B_1}_{1-uw_\lambda}\left(\mathbb{Z}_{B_1}\middle|\mathbb{X}_m,\mathbb{Y}_n\right)\right)>0\right)\leq\alpha\right\},$$

we deduce that

$$u^{B_2}_\alpha\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)\geq u^*=\frac{3\alpha}{4}$$

for $B_2 \geq \frac{8}{\alpha^2}\ln\left(\frac{2}{\beta}\right)$ when the event $\mathcal{A}$ holds.

Under the event $\mathcal{A}$, after performing B_3 steps of the bisection method, we have

$$\begin{aligned}
\widehat{u}^{B_{2:3}}_\alpha\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)&\geq u^{B_2}_\alpha\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)-\frac{\min_{\lambda\in\Lambda}w_\lambda^{-1}}{2^{B_3}}\\
&\geq\frac{3\alpha}{4}-\frac{\min_{\lambda\in\Lambda}w_\lambda^{-1}}{2^{B_3}}\\
&\geq\frac{\alpha}{2}
\end{aligned}$$

for $B_3 \geq \log_2\Big(\frac{4}{\alpha}\min_{\lambda\in\Lambda} w_\lambda^{-1}\Big)$.

We are interested in upper bounding the probability of type II error $\mathbb{P}_{p\times q\times r\times r}(\mathcal{B})$ for the event $\mathcal{B} \coloneqq \Big\{\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1},\mathbb{Z}_{B_2})=0\Big\}$. We have

$$\begin{aligned}\mathbb{P}_{p\times q\times r\times r}(\mathcal{B}) &= \mathbb{P}_{p\times q\times r\times r}\big(\mathcal{B}\big|\mathcal{A}\big)\,\mathbb{P}_{p\times q\times r\times r}(\mathcal{A}) + \mathbb{P}_{p\times q\times r\times r}\big(\mathcal{B}\big|\mathcal{A}^c\big)\,\mathbb{P}_{p\times q\times r\times r}(\mathcal{A}^c)\\ &\leq \mathbb{P}_{p\times q\times r\times r}\big(\mathcal{B}\big|\mathcal{A}\big) + \frac{\beta}{2}\end{aligned}$$

where

$$\begin{aligned}&\mathbb{P}_{p\times q\times r\times r}\big(\mathcal{B}\big|\mathcal{A}\big)\\ &= \mathbb{P}_{p\times q\times r\times r}\Big(\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1},\mathbb{Z}_{B_2})=0\,\Big|\,\mathcal{A}\Big)\\ &= \mathbb{P}_{p\times q\times r\times r}\left(\bigcap_{\lambda\in\Lambda}\left\{\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\leq \widehat{q}^{\lambda,B_1}_{1-\widehat{u}_\alpha^{B_{2:3}}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)\right\}\,\middle|\,\mathcal{A}\right)\\ &\leq \min_{\lambda\in\Lambda}\ \mathbb{P}_{p\times q\times r\times r}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\leq \widehat{q}^{\lambda,B_1}_{1-\widehat{u}_\alpha^{B_{2:3}}\left(\mathbb{Z}_{B_2}\middle|\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1}\right)w_\lambda}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)\,\middle|\,\mathcal{A}\right)\\ &\leq \min_{\lambda\in\Lambda}\ \mathbb{P}_{p\times q\times r}\Big(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)\leq \widehat{q}^{\lambda,B_1}_{1-\alpha w_\lambda/2}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)\Big)\\ &= \min_{\lambda\in\Lambda}\ \mathbb{P}_{p\times q\times r}\Big(\Delta^{\lambda,B_1}_{\alpha w_\lambda/2}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)=0\Big),\end{aligned}$$

we deduce that

$$\mathbb{P}_{p\times q\times r\times r}\big(\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1},\mathbb{Z}_{B_2})=0\big)\leq\frac{\beta}{2}+\min_{\lambda\in\Lambda}\ \mathbb{P}_{p\times q\times r}\Big(\Delta^{\lambda,B_1}_{\alpha w_\lambda/2}\big(\mathbb{Z}_{B_1}\big|\mathbb{X}_m,\mathbb{Y}_n\big)=0\Big).\tag{25}$$

In order to upper bound $\mathbb{P}_{p\times q\times r\times r}\Big(\Delta_\alpha^{\Lambda^w,B_{1:3}}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1},\mathbb{Z}_{B_2})=0\Big)$ by β it is sufficient to upper bound $\min_{\lambda\in\Lambda}\ \mathbb{P}_{p\times q\times r}\Big(\Delta^{\lambda,B_1}_{\alpha w_\lambda/2}(\mathbb{X}_m,\mathbb{Y}_n,\mathbb{Z}_{B_1})=0\Big)$ by $\frac{\beta}{2}$. By definition of uniform separation rates, it follows that

$$\rho\big(\Delta_\alpha^{\Lambda^w,B_{1:3}},\mathcal{S}^s_d(R),\beta,M\big)^2\leq 4\min_{\lambda\in\Lambda}\ \rho\Big(\Delta^{\lambda,B_1}_{\alpha w_\lambda/2},\mathcal{S}^s_d(R),\beta,M\Big)^2.$$

For each $\lambda\in\lambda$, since $\big(1-\alpha w_\lambda/2\big)\alpha w_\lambda/2\leq(1-\alpha)\alpha$ as $\alpha\in(0,e^{-1})$ and $w_\lambda\leq 1$, we have

$$\begin{aligned}B_1 &\geq \Big(\max\nolimits_{\lambda\in\Lambda} w_\lambda^{-2}\Big)\frac{12}{\alpha^2}\left(\log\left(\frac{8}{\beta}\right)+\alpha(1-\alpha)\right)\\ &\geq \frac{3}{(\alpha w_\lambda/2)^2}\left(\log\left(\frac{8}{\beta}\right)+\frac{\alpha w_\lambda}{2}\left(1-\frac{\alpha w_\lambda}{2}\right)\right),\end{aligned}$$

so we can apply Theorem 6 to the tests $\left(\Delta_{\alpha w_\lambda/2}^{\lambda,B_1}\right)_{\lambda\in\Lambda}$ to obtain

$$\begin{aligned}\rho\big(\Delta_\alpha^{\Lambda^w,B_{1:3}},\mathcal{S}_d^s(R),\beta,M\big)^2 &\le 4\min_{\lambda\in\Lambda}\ \rho\Big(\Delta_{\alpha w_\lambda/2}^{\lambda,B_1},\mathcal{S}_d^s(R),\beta,M\Big)^2\\ &\le 4C_4(M,d,s,R,\beta)\min_{\lambda\in\Lambda}\left(\sum_{i=1}^d\lambda_i^{2s}+\frac{\ln\left(\frac{2}{\alpha w_\lambda}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\\ &\le C_6(M,d,s,R,\beta)\min_{\lambda\in\Lambda}\left(\sum_{i=1}^d\lambda_i^{2s}+\frac{\ln\left(\frac{1}{\alpha w_\lambda}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\end{aligned}$$

for $C_6(M,d,s,R,\beta)\coloneqq 8C_4(M,d,s,R,\beta)$ where $C_4(M,d,s,R,\beta)$ is the constant from Theorem 6, and where we used the fact that

$$\ln\left(\frac{2}{\alpha w_\lambda}\right)=\ln(2)+\ln\left(\frac{1}{\alpha w_\lambda}\right)\le\big(\ln(2)+1\big)\ln\left(\frac{1}{\alpha w_\lambda}\right)\le 2\ln\left(\frac{1}{\alpha w_\lambda}\right)$$

as $\ln\left(\frac{1}{\alpha w_\lambda}\right)\ge\ln\left(\frac{1}{\alpha}\right)>1$ since $\alpha\in(0,e^{-1})$.

E.10 Proof of Corollary 10

First, note that we indeed have

$$\sum_{\lambda\in\Lambda}w_\lambda<\frac{6}{\pi^2}\sum_{\ell=1}^\infty\frac{1}{\ell^2}=1$$

and also that for all $\lambda=(2^{-\ell},\dots,2^{-\ell})\in\Lambda$ we have $\lambda_1\cdots\lambda_d=2^{-d\ell}<1$ as $\ell,d\in\mathbb{N}\setminus\{0\}$. Let $\lambda^*=(2^{-\ell^*},\dots,2^{-\ell^*})\in\Lambda$ where

$$\ell^*\coloneqq\left\lceil\frac{2}{4s+d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)\right\rceil\le\left\lceil\frac{2}{d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)\right\rceil.$$

Since $\min_{\lambda\in\Lambda}w_\lambda^{-1}=\frac{6}{\pi^2}$, we have $B_3\ge\log_2\left(\frac{4}{\alpha}\min_{\lambda\in\Lambda}w_\lambda^{-1}\right)$, so we can apply Theorem 9 to get

$$\begin{aligned}\rho\big(\Delta_\alpha^{\Lambda^w,B_{1:3}},\mathcal{S}_d^s(R),\beta,M\big)^2 &\le C_6(M,d,s,R,\beta)\min_{\lambda\in\Lambda}\left(\sum_{i=1}^d\lambda_i^{2s}+\frac{\ln\left(\frac{1}{\alpha}\right)+\ln\left(\frac{1}{w_\lambda}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right)\\ &\le C_6(M,d,s,R,\beta)\left(\sum_{i=1}^d(\lambda_i^*)^{2s}+\frac{\ln\left(\frac{1}{\alpha}\right)+\ln\left(\frac{1}{w_{\lambda^*}}\right)}{(m+n)\sqrt{\lambda_1^*\cdots\lambda_d^*}}\right).\end{aligned}$$

Note that $\ell^*\le\frac{2}{4s+d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)+1$ which gives $\lambda_i^*=2^{-\ell^*}\ge 2^{-1}\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-2/(4s+d)}$ for $i=1,\dots,d$. We get $\sqrt{\lambda_1^*\cdots\lambda_d^*}\ge 2^{-\frac{d}{2}}\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-d/(4s+d)}$ and so

$$\frac{1}{\sqrt{\lambda_1^*\cdots\lambda_d^*}}\le 2^{\frac{d}{2}}\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{d/(4s+d)}.$$

Note also that

$$\begin{aligned}\ell^* &\leq \frac{2}{4s+d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)+1\\ &\leq \frac{2}{4s+d}\log_2(m+n)+1\\ &\leq \left(\frac{2}{d\ln(2)}+1\right)\ln(m+n)\\ &< 4\ln(m+n)\end{aligned}$$

as $\ln(\ln(m+n))>1$ and $\ln(m+n)>1$. We get

$$\begin{aligned}\ln\left(\frac{1}{w_{\lambda^*}}\right) &= 2\ln(\ell^*)+\ln\left(\frac{\pi^2}{6}\right)\\ &\leq 2\ln(4\ln(m+n))+\ln\left(\frac{\pi^2}{6}\right)\\ &\leq \left(2\ln(4)+1+\ln\left(\frac{\pi^2}{6}\right)\right)\ln(\ln(m+n))\\ &< 5\ln(\ln(m+n))\end{aligned}$$

as $\ln(\ln(m+n))>1$. Combining those upper bounds, we get

$$\begin{aligned}\frac{\ln\left(\frac{1}{\alpha}\right)+\ln\left(\frac{1}{w_{\lambda^*}}\right)}{(m+n)\sqrt{\lambda_1^*\dots\lambda_d^*}} &\leq \frac{1}{(m+n)\sqrt{\lambda_1^*\dots\lambda_d^*}}\left(\ln\left(\frac{1}{\alpha}\right)+5\ln(\ln(m+n))\right)\\ &\leq \left(\ln\left(\frac{1}{\alpha}\right)+5\right)\frac{\ln(\ln(m+n))}{m+n}\frac{1}{\sqrt{\lambda_1^*\dots\lambda_d^*}}\\ &\leq 2^{\frac{d}{2}}\left(\ln\left(\frac{1}{\alpha}\right)+5\right)\frac{\ln(\ln(m+n))}{m+n}\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{d/(4s+d)}\\ &= 2^{\frac{d}{2}}\left(\ln\left(\frac{1}{\alpha}\right)+5\right)\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-4s/(4s+d)}\end{aligned}$$

as $\ln(\ln(m+n))>1$. Note also that

$$\ell^* \geq \frac{2}{4s+d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)$$

giving

$$(\lambda_i^*)^{2s} = (2^{-\ell^*})^{2s} \leq \left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-4s/(4s+d)}$$

for $i=1,\dots,d$. Hence, we get

$$\sum_{i=1}^{d}(\lambda_i^*)^{2s} \leq d\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-4s/(4s+d)}.$$

We obtain

$$\rho\big(\Delta_\alpha^{\Lambda^w,B_{1:3}},\mathcal{S}_d^s(R),\beta,M\big)^2 \leq C_6(M,d,s,R,\beta)\left(\sum_{i=1}^d(\lambda_i^*)^{2s}+\frac{\ln\big(\frac{1}{\alpha}\big)+\ln\Big(\frac{1}{w_{\lambda^*}}\Big)}{(m+n)\sqrt{\lambda_1^*\cdots\lambda_d^*}}\right)$$
$$\leq C_7(M,d,s,R,\alpha,\beta)^2\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-4s/(4s+d)}$$

where $C_7(M,d,s,R,\alpha,\beta) \coloneqq \sqrt{C_6(M,d,s,R,\beta)\max\Big\{d,2^{\frac{d}{2}}\big(\ln\big(\frac{1}{\alpha}\big)+5\big)\Big\}}$. We conclude that

$$\rho\big(\Delta_\alpha^{\Lambda^w,B_{1:3}},\mathcal{S}_d^s(R),\beta,M\big) \leq C_7(M,d,s,R,\alpha,\beta)\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-2s/(4s+d)}.$$

Hence, the test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ is optimal in the minimax sense up to an iterated logarithmic term. Since it does not depend on the unknown parameters s and R, our aggregated MMDAgg test $\Delta_\alpha^{\Lambda^w,B_{1:3}}$ is minimax adaptive over the Sobolev balls $\big\{\mathcal{S}_d^s(R) : s>0, R>0\big\}$.

References

M. Albert, B. Laurent, A. Marrel, and A. Meynaoui. Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879, 2022.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 1 (8(5):577–606), 2002.

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.

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

P. L. Chebyshev. Oeuvres. *Commissionaires de l'Académie Impériale des Sciences*, 1, 1899.

S. X. Chen and Y.-L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615. PMLR, 2016.

V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer Science & Business Media, 1999.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

H. S. Fisher, B. B. Wong, and G. G. Rosenthal. Alteration of the chemical environment disrupts communication in a freshwater fish. *Proceedings of the Royal Society B: Biological Sciences*, 273(1591):1187–1193, 2006.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

M. Fromont, B. Laurent, M. Lerasle, and P. Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Conference on Learning Theory*, volume 23 of *Journal of Machine Learning Research Proceedings*, 2012.

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.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 1, pages 489–496, 2008.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*. Springer, 2005.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems*, pages 513–520, Cambridge, MA, 2007. MIT Press.

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, 2012a.

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*, volume 1, pages 1205–1213, 2012b.

W. Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in Statistics*, pages 308–334. Springer, 1992.

L. Horváth, P. Kokoszka, and R. Reeder. Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):103–122, 2013.

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

Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. *Journal of Soviet Mathematics*, 1(44:466–476), 1993a.

Y. I. Ingster. Minimax testing of the hypothesis of independence for ellipsoids in l_p. *Zapiski Nauchnykh Seminarov POMI*, 1(207:77–97), 1993b.

W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, volume 29, pages 181–189, 2016.

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

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

J. M. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet. Learning kernel tests without data splitting. In *Advances in Neural Information Processing Systems 33*, pages 6245–6255. Curran Associates, Inc., 2020.

J. M. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pages 1403–1419. PMLR, 2022a.

J. M. Kübler, V. Stimper, S. Buchholz, K. Muandet, and B. Schölkopf. AutoML two-sample test. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022b.

Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database. AT&T Labs, 2010. URL `http://yann.lecun.com/exdb/mnist/`.

J. Lee. *U-statistics: Theory and Practice*. Citeseer, 1990.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

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

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*, 2020.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284. PMLR, 2016.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

F. J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

R. R. Miles, R. F. Roberts, A. R. Putnam, and W. L. Roberts. Comparison of serum and heparinized plasma samples for measurement of chemistry analytes. *Clinical Chemistry*, 50(9):1704–1706, 2004.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1:429–443, 1997.

S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

S. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR, 2015.

J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005a.

J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005b.

A. Schrab, B. Guedj, and A. Gretton. KSD aggregated goodness-of-fit test. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022a.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient aggregated kernel tests using incomplete U-statistics. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022b.

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Student. The probable error of a mean. *Biometrika*, 1(1):1–25, 1908.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

I. Tolstikhin, B. a. K. Sriperumbudur, and B. Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

P. Vermeesch. Multi-sample comparison of detrital age distributions. *Chemical Geology*, 341:140–146, 2013.

G. Wynne and A. B. Duncan. A kernel two-sample test for functional data. *Journal of Machine Learning Research*, 23(73):1–51, 2022.

G. Wynne and S. Nagy. Statistical depth meets machine learning: Kernel mean embeddings and depth in functional data analysis. *arXiv preprint arXiv:2105.12778*, 2021.

M. Yamada, D. Wu, Y. H. Tsai, H. Ohta, R. Salakhutdinov, I. Takeuchi, and K. Fukumizu. Post selection inference with incomplete maximum mean discrepancy estimator. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.