Efficient Aggregated Kernel Tests using Incomplete U-statistics

Antonin Schrab
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit
University College London & Inria London
a.schrab@ucl.ac.uk

Ilmun Kim
Department of Statistics & Data Science
Department of Applied Statistics
Yonsei University
ilmun@yonsei.ac.kr

Benjamin Guedj
Centre for Artificial Intelligence
University College London & Inria London
b.guedj@ucl.ac.uk

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

Abstract

We propose a series of computationally efficient, nonparametric tests for the two-sample, independence and goodness-of-fit problems, using the Maximum Mean Discrepancy (MMD), Hilbert Schmidt Independence Criterion (HSIC), and Kernel Stein Discrepancy (KSD), respectively. Our test statistics are incomplete U-statistics, with a computational cost that interpolates between linear time in the number of samples, and quadratic time, as associated with classical U-statistic tests. The three proposed tests aggregate over several kernel bandwidths to detect departures from the null on various scales: we call the resulting tests MMDAggInc, HSICAggInc and KSDAggInc. This procedure provides a solution to the fundamental kernel selection problem as we can aggregate a large number of kernels with several bandwidths without incurring a significant loss of test power. For the test thresholds, we derive a quantile bound for wild bootstrapped incomplete U-statistics, which is of independent interest. We derive non-asymptotic uniform separation rates for MMDAggInc and HSICAggInc, and quantify exactly the trade-off between computational efficiency and the attainable rates: this result is novel for tests based on incomplete U-statistics, to our knowledge. We further show that in the quadratic-time case, the wild bootstrap incurs no penalty to test power over more widespread permutation-based approaches, since both attain the same minimax optimal rates (which in turn match the rates that use oracle quantiles). We support our claims with numerical experiments on the trade-off between computational efficiency and test power.

1 Introduction

Nonparametric hypothesis testing is a fundamental field of statistics, and is widely used by the machine learning community and practitioners in numerous other fields, due to the increasing availability of huge amounts of data. When dealing with large-scale datasets, computational cost can quickly emerge as a major issue which might prevent from using expensive tests in practice; constructing efficient tests is therefore crucial for their real-world applications. In this paper, we construct kernel-based aggregated tests using incomplete U-statistics (Blom, 1976) for the **two-sample**, **independence** and **goodness-of-fit** problems (which we detail in Section 2). The quadratic-time aggregation procedure has been shown to result in powerful tests (Fromont et al., 2012; Fromont et al., 2013; Albert et al.,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2022; Schrab et al., 2021; Schrab et al., 2022), and we propose efficient variants of these well-studied tests, with computational cost interpolating from the classical quadratic-time regime to the linear-time one.

Related work: aggregated tests. Kernel selection (or kernel bandwidth selection) is a fundamental problem in nonparametric hypothesis testing because it has a major influence on test power. Motivated by this problem, non-asymptotic aggregated tests, which combine tests with different kernel bandwidths, have been proposed for the two-sample (Fromont et al., 2012, 2013; Kim et al., 2022; Schrab et al., 2021), independence (Albert et al., 2022; Kim et al., 2022), and goodness-of-fit (Schrab et al., 2022) testing frameworks. Li and Yuan (2019) and Balasubramanian et al. (2021) construct similar aggregated tests for these three problems, with the difference that they work in the asymptotic regime. All the mentioned works study aggregated tests in terms of uniform separation rates (Baraud, 2002). Those rates depend on the sample size and satisfy the following property: if the L^2-norm difference between the densities is greater than the uniform separation rate, then the test is guaranteed to have high power. All aggregated kernel-based tests in the existing literature have been studied using estimators which are U-statistics (Hoeffding, 1992) with tests running in quadratic time.

Related work: efficient kernel tests. Several linear-time kernel tests have been proposed for those three testing frameworks. Those include tests using classical linear-time estimators with median bandwidth (Gretton et al., 2012a; Liu et al., 2016) or selecting an optimal bandwidth on held-out data to maximize power (Gretton et al., 2012b), tests using eigenspectrum approximation (Gretton et al., 2009), tests using post-selection inference for adaptive kernel selection, also using incomplete U-statistics (Yamada et al., 2018, 2019; Lim et al., 2019, 2020; Kübler et al., 2020; Freidling et al., 2021), tests which use a Nyström approximation of the asymptotic null distribution (Zhang et al., 2018; Cherfaoui et al., 2022), random Fourier features tests (Zhang et al., 2018; Zhao and Meng, 2015; Chwialkowski et al., 2015), tests based on random feature Stein discrepancies (Huggins and Mackey, 2018), and the adaptive tests which use features selected on held-out data to maximize power (Jitkrittum et al., 2016, 2017a,b), as well as tests using neural networks to learn a discrepancy (Grathwohl et al., 2020). We also point out the very relevant works of Kübler et al. (2022) on a quadratic-time test, and of Ho and Shieh (2006), Zaremba et al. (2013) and Zhang et al. (2018) on the use of block U-statistics with complexity $\mathcal{O}(N^{1.5})$ for block size $\sqrt{N}$ where N is the sample size.

Contributions and outline. In Section 2, we present the three testing problems with their associated well-known quadratic-time kernel-based estimators (MMD, HSIC, KSD) which are U-statistics. We introduce three associated incomplete U-statistics estimators, which can be computed efficiently, in Section 3. We then provide quantile and variance bounds for generic incomplete U-statistics using a wild bootstrap, in Section 4. We study the level and power guarantees at every finite sample sizes for our efficient tests using incomplete U-statistics for a fixed kernel bandwidth, in Section 5. In particular, we obtain non-asymptotic uniform separation rates for the two-sample and independence tests over a Sobolev ball, and show that these rates are minimax optimal up to the cost incurred for efficiency of the test. In Section 6, we propose our efficient aggregated tests which combine tests with multiple kernel bandwidths. We prove that the proposed tests are adaptive over Sobolev balls and achieve the same uniform separation rate (up to an iterated logarithmic term) as the tests with optimal bandwidths. As a result of our analysis, we have shown minimax optimality over Sobolev balls of the quadratic-time tests using quantiles estimated with a wild bootstrap. Whether this optimality result also holds for tests using the more general permutation-based procedure to approximate HSIC quantiles, was an open problem formulated by Kim et al. (2022), we prove that it indeed holds in Section 7. We close the paper with numerical experiments in Section 8, where we observe that linear-time MMDAggInc, HSICAggInc and KSDAggInc retain high power and outperform other state-of-the-art linear-time kernel tests. Our implementation of the tests and code for reproducibility of the experiments are available online under the MIT License: `https://github.com/antoninschrab/agginc-paper`.

2 Background

Here we briefly describe our main problems of interest, comprising the two-sample, independence and goodness-of-fit problems. We approach these problems from a nonparametric point of view using the kernel-based statistics: MMD, HSIC, and KSD. We briefly introduce original forms of these statistics, which can be computed in quadratic time, and also discuss ways of calibrating tests proposed in the literature. The three quadratic-time expressions are presented in Appendix B.

Two-sample testing. In this problem, we are 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[1] p and q on $\mathbb{R}^d$. We assume we work with balanced sample sizes, that is[2] $\max(m,n) \lesssim \min(m,n)$. We are interested in testing the null hypothesis $\mathcal{H}_0 : p = q$ against the alternative $\mathcal{H}_1 : p \neq q$; that is, we want to know if the samples come from the same distribution. Gretton et al. (2012a) propose a non-parametric kernel test based on the *Maximum Mean Discrepancy* (MMD), a measure between probability distributions which uses a characteristic kernel k (Fukumizu et al., 2008; Sriperumbudur et al., 2011). It can be estimated using a quadratic-time estimator (Gretton et al., 2012a, Lemma 6) which, as noted by Kim et al. (2022), can be expressed as a two-sample U-statistic (both of second order) (Hoeffding, 1992),

$$\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n) = \frac{1}{\left|\mathbf{i}_2^m\right|\left|\mathbf{i}_2^n\right|} \sum_{(i,i')\in\mathbf{i}_2^m} \sum_{(j,j')\in\mathbf{i}_2^n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_j, Y_{j'}), \tag{1}$$

where $\mathbf{i}_a^b$ denotes the set of all a-tuples drawn without replacement from $\{1,\dots,b\}$ so that $\left|\mathbf{i}_a^b\right| = b\cdots(b-a+1)$, and where, for $x_1, x_2, y_1, y_2 \in \mathbb{R}^d$, we let

$$h_k^{\mathrm{MMD}}(x_1, x_2; y_1, y_2) := k(x_1, x_2) - k(x_1, y_2) - k(x_2, y_1) + k(y_1, y_2). \tag{2}$$

Independence testing. In this problem, we have access to i.i.d. pairs of samples $\mathbb{Z}_N := \big(Z_i\big)_{1\leq i\leq N} = \big((X_i, Y_i)\big)_{1\leq i\leq N}$ with joint probability density p_{xy} on $\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$ and marginals p_x on $\mathbb{R}^{d_x}$ and p_y on $\mathbb{R}^{d_y}$. We are interested in testing $\mathcal{H}_0 : p_{xy} = p_x \otimes p_y$ against $\mathcal{H}_1 : p_{xy} \neq p_x \otimes p_y$; that is, we want to know if two components of the pairs of samples are independent or dependent. Gretton et al. (2005, 2008) propose a non-parametric kernel test based on the *Hilbert Schmidt Independence Criterion* (HSIC). It can be estimated using the quadratic-time estimator proposed by Song et al. (2012, Equation 5) which is a fourth-order one-sample U-statistic

$$\widehat{\mathrm{HSIC}}_{k,\ell}(\mathbb{Z}_N) = \frac{1}{\left|\mathbf{i}_4^N\right|} \sum_{(i,j,r,s)\in\mathbf{i}_4^N} h_{k,\ell}^{\mathrm{HSIC}}(Z_i, Z_j, Z_r, Z_s) \tag{3}$$

for characteristic kernels k on $\mathbb{R}^{d_x}$ and ℓ on $\mathbb{R}^{d_y}$ (Gretton, 2015), and where for $z_a = (x_a, y_a) \in \mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$, $a = 1,\dots,4$, we let

$$h_{k,\ell}^{\mathrm{HSIC}}(z_1, z_2, z_3, z_4) := \frac{1}{4} h_k^{\mathrm{MMD}}(x_1, x_2; x_3, x_4) h_\ell^{\mathrm{MMD}}(y_1, y_2; y_3, y_4). \tag{4}$$

Goodness-of-fit testing. For this problem, we are given a model density p on $\mathbb{R}^d$ and i.i.d. samples $\mathbb{Z}_N := (Z_i)_{1\leq i\leq N}$ drawn from a density q on $\mathbb{R}^d$. The aim is again to test $\mathcal{H}_0 : p = q$ against $\mathcal{H}_1 : p \neq q$; that is, we want to know if the samples have been drawn from the model. Chwialkowski et al. (2016) and Liu et al. (2016) both construct a non-parametric goodness-of-fit test using the *Kernel Stein Discrepancy* (KSD). A quadratic-time KSD estimator can be computed as the second-order one-sample U-statistic,

$$\widehat{\mathrm{KSD}}_{p,k}^2(\mathbb{Z}_N) := \frac{1}{\left|\mathbf{i}_2^N\right|} \sum_{(i,j)\in\mathbf{i}_2^N} h_{k,p}^{\mathrm{KSD}}(Z_i, Z_j), \tag{5}$$

where the *Stein kernel* $h_{p,k} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{aligned} h_{k,p}^{\mathrm{KSD}}(x,y) := {} & \left(\nabla \log p(x)^\top \nabla \log p(y)\right) k(x,y) + \nabla \log p(y)^\top \nabla_x k(x,y) \\ & + \nabla \log p(x)^\top \nabla_y k(x,y) + \sum_{i=1}^d \frac{\partial}{\partial x_i \partial y_i} k(x,y). \end{aligned} \tag{6}$$

In order to guarantee consistency of the Stein goodness-of-fit test (Chwialkowski et al., 2016, Theorem 2.2), we assume that the kernel k is C_0-universal (Carmeli et al., 2010, Definition 4.1) and that

$$\mathbb{E}_q\Big[h_{k,p}^{\mathrm{KSD}}(x,x)\Big] < \infty \qquad \text{and} \qquad \mathbb{E}_q\Big[\left\|\nabla \log\left(\frac{p(x)}{q(x)}\right)\right\|_2^2\Big] < \infty. \tag{7}$$

[1] All probability density functions in this paper are with respect to the Lebesgue measure.

[2] We use the notation $a \lesssim b$ when there exists a constant $C > 0$ such that $a \leq Cb$. We similarly use the notation $\gtrsim$. We write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. We also use the convention that all constants are generically denoted by C, even though they might be different.

Quantile estimation. Multiple strategies have been proposed to estimate the quantiles of test statistics under the null for these three tests. We primarily focus on the wild bootstrap approach (Chwialkowski et al., 2014), though our results also hold using a parametric bootstrap for the goodness-of-fit setting (Schrab et al., 2022). In Section 7, we show that the same uniform separation rates can be derived for HSIC quadratic-time tests using permutations instead of a wild bootstrap.

More details on MMD, HSIC, KSD, and on quantile estimation are provided in Appendix B.

3 Incomplete U-statistics for MMD, HSIC and KSD

As presented above, the quadratic-time statistics for the two-sample (MMD), independence (HSIC) and goodness-of-fit (KSD) problems can be rewritten as U-statistics with kernels h_k^{MMD}, $h_{k,\ell}^{\mathrm{HSIC}}$ and h_k^{KSD}, respectively. The computational cost of tests based on these U-statistics grows quadratically with the sample size. When working with very large sample sizes, as it is often the case in real-world uses of those tests, this quadratic cost can become very problematic, and faster alternative tests are better adapted to this 'big data' setting. Multiple linear-time kernel tests have been proposed in the three testing frameworks (see Section 1 for details). We construct computationally efficient variants of the aggregated kernel tests proposed by Fromont et al. (2013), Albert et al. (2022), Kim et al. (2022), and Schrab et al. (2021, 2022) for the three settings, with the aim of retaining the significant power advantages of the aggregation procedure observed for quadratic-time tests. To this end, we propose to replace the quadratic-time U-statistics presented in Equations (1), (3) and (5) with second order incomplete U-statistics (Blom, 1976; Janson, 1984; Lee, 1990),

$$\overline{\mathrm{MMD}}_k^2\big(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{D}_N\big) \coloneqq \frac{1}{\big|\mathcal{D}_N\big|} \sum_{(i,j)\in\mathcal{D}_N} h_k^{\mathrm{MMD}}(X_i, X_j; Y_i, Y_j), \tag{8}$$

$$\overline{\mathrm{HSIC}}_{k,\ell}\big(\mathbb{Z}_N; \mathcal{D}_{\lfloor N/2 \rfloor}\big) \coloneqq \frac{1}{\big|\mathcal{D}_{\lfloor N/2 \rfloor}\big|} \sum_{(i,j)\in\mathcal{D}_{\lfloor N/2 \rfloor}} h_{k,\ell}^{\mathrm{HSIC}}\left(Z_i, Z_j, Z_{i+\lfloor N/2 \rfloor}, Z_{j+\lfloor N/2 \rfloor}\right), \tag{9}$$

$$\overline{\mathrm{KSD}}_{p,k}^2\big(\mathbb{Z}_N; \mathcal{D}_N\big) \coloneqq \frac{1}{\big|\mathcal{D}_N\big|} \sum_{(i,j)\in\mathcal{D}_N} h_{k,p}^{\mathrm{KSD}}(Z_i, Z_j), \tag{10}$$

where for the two-sample problem we let $N \coloneqq \min(m, n)$, and where the *design* $\mathcal{D}_b$ is a subset of $\mathbf{i}_2^b$ (the set of all 2-tuples drawn without replacement from $\{1, \dots, b\}$). Note that $\mathcal{D}_{\lfloor N/2 \rfloor} \subseteq \mathbf{i}_2^{N/2} \subset \mathbf{i}_2^N$. The design can be deterministic. For example, for the two-sample problem with equal even sample sizes $m = n = N$, the deterministic design $\mathcal{D}_N = \{(2a-1, 2a) : a = 1, \dots, N/2\}$ corresponds to the MMD linear-time estimator proposed by Gretton et al. (2012a, Lemma 14). For fixed design size, the elements of the design can also be chosen at random without replacement, in which case the estimators in Equations (8) to (10) become random quantities given the data. For generality purposes, the results presented in this paper hold for both deterministic and random (without replacement) design choices while we focus on the deterministic design in our experiments. By fixing the design sizes in Equations (8) to (10) to be, for example,

$$\big|\mathcal{D}_N\big| = \big|\mathcal{D}_{\lfloor N/2 \rfloor}\big| = cN \tag{11}$$

for some small constant $c \in \mathbb{N} \setminus \{0\}$, we obtain incomplete U-statistics which can be computed in linear time. Note that by pairing the samples $Z_i \coloneqq (X_i, Y_i)$, $i = 1, \dots, N$ for the MMD case and $\widetilde{Z}_i \coloneqq \big(Z_i, Z_{i+\lfloor N/2 \rfloor}\big)$, $i = 1, \dots, \lfloor N/2 \rfloor$ for the HSIC case, we observe that all three incomplete U-statistics of second order have the same form, with only the kernel functions and the design differing. The motivation for defining the estimators in Equations (8) and (9) as incomplete U-statistics of order 2 (rather than of higher order) derives from the reasoning of Kim et al. (2022, Section 6) using permuted complete U-statistics for the two-sample and independence problems (see Appendix E.1).

4 Quantile and variance bounds for incomplete U-statistics

Here we derive upper quantile and variance bounds for a second order incomplete degenerate U-statistic with a generic degenerate kernel h, for some design $\mathcal{D} \subseteq \mathbf{i}_2^N$, defined as

$$\overline{U}\big(\mathbb{Z}_N; \mathcal{D}\big) \coloneqq \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} h(Z_i, Z_j).$$

We will use these results to bound the quantiles and variances of our three test statistics for our hypothesis tests in Section 5. The derived bounds are of independent interest.

In the following lemma, building on the results of Lee (1990), we directly derive an upper bound on the variance of the incomplete U-statistic in terms of the sample size N and of the design size $|\mathcal{D}|$.

Lemma 1. *The variance of the incomplete U-statistic can be upper bounded in terms of the quantities $\sigma_1^2 := \mathrm{var}\big(\mathbb{E}\big[h(Z,Z')\big|Z'\big]\big)$ and $\sigma_2^2 := \mathrm{var}(h(Z,Z'))$ with different bounds depending on the design choice. For deterministic (LHS) or random (RHS) design $\mathcal{D}$ and sample size N, we have*

$$\mathrm{var}\big(\overline{U}\big) \lesssim \frac{N}{|\mathcal{D}|}\sigma_1^2 + \frac{1}{|\mathcal{D}|}\sigma_2^2 \qquad \textit{and} \qquad \mathrm{var}\big(\overline{U}\big) \lesssim \frac{1}{N}\sigma_1^2 + \frac{1}{|\mathcal{D}|}\sigma_2^2.$$

The proof of Lemma 1 is deferred to Appendix F.2. We emphasize the fact that this variance bound also holds for random design with replacement, as considered by Blom (1976) and Lee (1990). For random design, we observe that if $|\mathcal{D}| \asymp N^2$ then the bound is $\sigma_1^2/N + \sigma_2^2/N^2$ which is the variance bound of the complete U-statistic (Albert et al., 2022, Lemma 10). If $N \lesssim |\mathcal{D}| \lesssim N^2$, the variance bound is $\sigma_1^2/N + \sigma_2^2/|\mathcal{D}|$, and if $|\mathcal{D}| \lesssim N$ it is $\sigma_2^2/|\mathcal{D}|$ since $\sigma_1^2 \leq \sigma_2^2/2$ (Blom, 1976, Equation 2.1).

Kim et al. (2022) develop exponential concentration bounds for permuted complete U-statistics, and Clémençon et al. (2013) study the uniform approximation of U-statistics by incomplete U-statistics. To the best of our knowledge, no quantile bounds have yet been obtained for incomplete U-statistics in the literature. While permutations are well-suited for complete U-statistics (Kim et al., 2022), using them with incomplete U-statistics results in having to compute new kernel values, and this comes at an extra computational cost we would like to avoid. Restricting the set of permutations to those for which the kernel values have already been computed for the original incomplete U-statistic corresponds exactly to using a wild bootstrap (Schrab et al., 2021, Appendix B). Hence, we consider the wild bootstrapped second order incomplete U-statistic

$$\overline{U}^{\epsilon}\big(\mathbb{Z}_N;\mathcal{D}\big) := \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}} \epsilon_i\epsilon_j h(Z_i,Z_j) \tag{12}$$

for i.i.d. Rademacher random variables $\epsilon_1,\ldots,\epsilon_N$ with values in $\{-1,1\}$, for which we derive an exponential concentration bound (quantile bound). We note the in-depth work of Chwialkowski et al. (2014) on the wild bootstrap procedure for kernel tests with applications to quadratic-time MMD and HSIC tests. We now provide exponential tail bounds for wild bootstrapped incomplete U-statistics.

Lemma 2. *There exists some constant $C > 0$ such that, for every $t \geq 0$, we have*

$$\mathbb{P}_{\epsilon}\Big(\big|\overline{U}^{\epsilon}\big| \geq t \,\big|\, \mathbb{Z}_N,\mathcal{D}\Big) \leq 2\exp\left(-C\frac{t}{A_{\mathrm{inc}}}\right) \leq 2\exp\left(-C\frac{t}{A}\right)$$

where $A_{\mathrm{inc}}^2 := |\mathcal{D}|^{-2}\sum_{(i,j)\in\mathcal{D}} h(Z_i,Z_j)^2$ and $A^2 := |\mathcal{D}|^{-2}\sum_{(i,j)\in\mathbf{i}_2^N} h(Z_i,Z_j)^2$.

Lemma 2 is proved in Appendix F.3. While the second bound in Lemma 2 is less tight, it has the benefit of not depending on the choice of design $\mathcal{D}$ but only on the design size $|\mathcal{D}|$ which is usually fixed.

5 Efficient kernel tests using incomplete U-statistics

We now formally define the hypothesis tests obtained using the incomplete U-statistics with a wild bootstrap. This is done for fixed kernel bandwidths $\lambda \in (0,\infty)^{d_x}, \mu \in (0,\infty)^{d_y}$, for the kernels[3]

$$k_\lambda(x,y) := \prod_{i=1}^{d_x}\frac{1}{\lambda_i}K_i\left(\frac{x_i-y_i}{\lambda_i}\right), \qquad \ell_\mu(x,y) := \prod_{i=1}^{d_y}\frac{1}{\mu_i}L_i\left(\frac{x_i-y_i}{\mu_i}\right), \tag{13}$$

for characteristic kernels $(x,y)\mapsto K_i(x-y)$, $(x,y)\mapsto L_i(x-y)$ on $\mathbb{R}\times\mathbb{R}$ for functions $K_i, L_i \in L^1(\mathbb{R})\cap L^2(\mathbb{R})$ integrating to 1. We unify the notation for the three testing frameworks. For the two-sample and goodness-of-fit problems, we work only with k_λ and have $d = d_x$. For the independence

[3] Our results are presented for bandwidth selection, but they hold for more general kernel selection settings, as considered by Schrab et al. (2022). The results for the goodness-of-fit problem hold for a wider range of kernels including the IMQ (inverse multiquadric) kernel (Gorham and Mackey, 2017), as in Schrab et al. (2022).

problem, we work with the two kernels k_λ and ℓ_μ, and for ease of notation we let $d := d_x + d_y$ and $\lambda_{d_x+i} := \mu_i$ for $i = 1, \dots, d_y$. We also simply write $p := p_{xy}$ and $q := p_x \otimes p_y$. We let $\overline{U}_\lambda$ and h_λ denote either $\overline{\mathrm{MMD}}^2_{k_\lambda}$ and $h^{\mathrm{MMD}}_{k_\lambda}$, or $\overline{\mathrm{HSIC}}_{k_\lambda,\ell_\mu}$ and $h^{\mathrm{HSIC}}_{k_\lambda,\ell_\mu}$, or $\overline{\mathrm{KSD}}^2_{p,k_\lambda}$ and $h^{\mathrm{KSD}}_{k_\lambda,p}$, respectively. We denote the design size of the incomplete U-statistics in Equations (8) to (10) by

$$L := \left|\mathcal{D}_N\right| = \left|\mathcal{D}_{\lfloor N/2 \rfloor}\right|.$$

For the three testing frameworks, we estimate the quantiles of the test statistics by simulating the null hypothesis using a wild bootstrap, as done in the case of complete U-statistics by Fromont et al. (2012), Schrab et al. (2021) for the two-sample problem, and by Schrab et al. (2022) for the goodness-of-fit problem. This is done by considering the original test statistic $U_\lambda^{B_1+1} := \overline{U}_\lambda$ together with B_1 wild bootstrapped incomplete U-statistics $U_\lambda^1, \dots, U_\lambda^{B_1}$ computed as in Equation (12), and estimating the $(1-\alpha)$-quantile with a Monte Carlo approximation

$$\widehat{q}^{\lambda}_{1-\alpha} := \inf\left\{ t \in \mathbb{R} : 1 - \alpha \leq \frac{1}{B_1+1} \sum_{b=1}^{B_1+1} \mathbf{1}\big(U_\lambda^b \leq t\big) \right\} = U_\lambda^{\bullet \lceil B_1(1-\alpha) \rceil}, \tag{14}$$

where $U_\lambda^{\bullet 1} \leq \cdots \leq U_\lambda^{\bullet B_1+1}$ are the sorted elements $U_\lambda^1, \dots, U_\lambda^{B_1+1}$. The test Δ^λ_α is defined as rejecting the null if the original test statistic $\overline{U}_\lambda$ is greater than the estimated $(1-\alpha)$-quantile, that is,

$$\Delta^\lambda_\alpha(\mathbb{Z}_N) := \mathbf{1}\big(\overline{U}_\lambda(\mathbb{Z}_N) > \widehat{q}^{\lambda}_{1-\alpha}\big).$$

We show in Proposition 1 that the test Δ^λ_α has well-calibrated asymptotic level for goodness-of-fit testing, and well-calibrated non-asymptotic level for two-sample and independence testing. The proof of the latter non-asymptotic guarantee is based on the exchangeability of $U_\lambda^1, \dots, U_\lambda^{B_1+1}$ under the null hypothesis along with the result of Romano and Wolf (2005, Lemma 1). A similar proof strategy can be found in Fromont et al. (2012, Proposition 2), Albert et al. (2022, Proposition 1), and Schrab et al. (2021, Proposition 1). The exchangeability of wild bootstrapped incomplete U-statistics for independence testing does not follow directly from the mentioned works. We show this through an intriguing connection between the MMD kernel and the HSIC kernel (proof deferred to Appendix F.1).

Proposition 1. *The test Δ^λ_α has level $\alpha \in (0,1)$,* i.e., $\mathbb{P}_{\mathcal{H}_0}\big(\Delta^\lambda_\alpha(\mathbb{Z}_N) = 1\big) \leq \alpha$. *This holds non-asymptotically for the two-sample and independence cases, and asymptotically for goodness-of-fit.*[4]

Having established the validity of the test Δ^λ_α, we now study power guarantees for it in terms of the L^2-norm of the difference in densities $\|p - q\|_2$. In Theorem 1, we show for the three tests that, if $\|p - q\|_2$ exceeds some threshold, we can guarantee high test power. For the two-sample and independence problems, we derive a uniform separation rate (Baraud, 2002) over Sobolev balls

$$\mathcal{S}^s_d(R) := \left\{ f \in L^1\big(\mathbb{R}^d\big) \cap L^2\big(\mathbb{R}^d\big) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\widehat{f}(\xi)|^2 \mathrm{d}\xi \leq (2\pi)^d R^2 \right\}, \tag{15}$$

with radius $R > 0$ and smoothness parameter $s > 0$. This uniform separation rate is the smallest value of t such that for any alternative with $\|p - q\|_2 > t$ and[5] $p - q \in \mathcal{S}^s_d(R)$ the probability of type II error of Δ^λ_α can be controlled by $\beta \in (0,1)$. Before presenting Theorem 1, we need to introduce more notation unified over the three testing frameworks; we define the integral transform T_λ as

$$(T_\lambda f)(x) := \int_{\mathbb{R}^d} f(x) \mathcal{K}_\lambda(x, y) \, \mathrm{d}y \tag{16}$$

for $f \in L^2(\mathbb{R}^d)$, $x \in \mathbb{R}^d$, where $\mathcal{K}_\lambda := k_\lambda$ for the two-sample problem, $\mathcal{K}_\lambda := k_\lambda \otimes \ell_\mu$ for the independence problem, and $\mathcal{K}_\lambda := h^{\mathrm{KSD}}_{k_\lambda,p}$ for the goodness-of-fit problem. Note that, for the two-sample and independence testing frameworks, since $\mathcal{K}_\lambda$ is translation-invariant, the integral transform corresponds to a convolution. However, this is not true for the goodness-of-fit framework as $h^{\mathrm{KSD}}_{k_\lambda,p}$ is not translation-invariant. We are now in a position to present our main contribution in Theorem 1: we derive a power guarantee condition for our tests using incomplete U-statistics, and a uniform separation rate over Sobolev balls for the two-sample and independence settings.

[4] Level is non-asymptotic for the goodness-of-fit case when using a parametric bootstrap (Schrab et al., 2022). For the goodness-of-fit setting, we also recall that the further assumptions in Equation (7) need to be satisfied.

[5] We stress that we only assume $p - q \in \mathcal{S}^s_d(R)$ and not $p, q \in \mathcal{S}^s_d(R)$ as considered by Li and Yuan (2019). Viewing q as a perturbed version of p, we only require that the perturbation is smooth (*i.e.* lies in a Sobolev ball).

Theorem 1. *(i) Suppose that assumptions $\boldsymbol{\mathcal{A}_1}(\|p\|_\infty, \|q\|_\infty, \alpha, \beta, B_1)$ in Appendix A hold. Consider $\lambda \in (0,\infty)^d$ with $\lambda_1 \cdots \lambda_d \lesssim 1$ and $\sigma_{2,\lambda}^2 := \mathbb{E}\big[h_\lambda(Z,Z')^2\big]$. For sample size N and design size L, if there exists some $C > 0$ such that*

$$\|p-q\|_2^2 \;\geq\; \|(p-q) - T_\lambda(p-q)\|_2^2 + C\frac{N}{L}\frac{\ln(1/\alpha)}{\beta}\sigma_{2,\lambda},$$

then $\mathbb{P}_{\mathcal{H}_1}\big(\Delta_\alpha^\lambda(\mathbb{Z}_N) = 0\big) \leq \beta$ (type II error), where $\sigma_{2,\lambda} \lesssim 1/\sqrt{\lambda_1 \cdots \lambda_d}$ for MMD and HSIC.

(ii) Fix $R > 0$ and $s > 0$, and consider the bandwidths $\lambda_i^* := (N/L)^{2/(4s+d)}$ for $i = 1, \ldots, d$. For MMD and HSIC, the uniform separation rate of $\Delta_\alpha^{\lambda^*}$ over the Sobolev ball $\mathcal{S}_d^s(R)$ is (up to a constant)

$$(N/L)^{2s/(4s+d)}.$$

The proof of Theorem 1 relies on the variance and quantile bounds presented in Lemmas 1 and 2, and also uses results of Albert et al. (2022) and Schrab et al. (2021, 2022) on complete U-statistics. The details can be found in Appendix F.4. The power condition in Theorem 1 corresponds to a variance-bias decomposition; for large bandwidths the bias term (first term) dominates, while for small bandwidths the variance term (second term which also controls the quantile) dominates. While the power guarantees of Theorem 1 hold for any design (either deterministic or sampled uniformly without replacement) of fixed size L, the choice of design still influences the performance of the test in practice. The variance (but not its upper bound) depends on the choice of design; certain choices lead to minimum variance of the incomplete U-statistic (Lee, 1990, Section 4.3.2).

We recall that the minimax (i.e. optimal) rate over the Sobolev ball $\mathcal{S}_d^s(R)$ is $(1/N)^{2s/(4s+d)}$ for the two-sample (Li and Yuan, 2019, Theorem 5 (ii)) and independence (Albert et al., 2022, Theorem 4) problems. We highlight that the rate for our incomplete U-statistic test has the same dependence in the exponent as the minimax rate; that is $(N/L)^{2s/(4s+d)} = (1/N)^{2s/(4s+d)}\left(N^2/L\right)^{2s/(4s+d)}$ where $L \lesssim N^2$ is the design size and N is the sample size. We reach the following conclusions.

- If $L \asymp N^2$ then the test runs in quadratic time and we recover exactly the minimax rate.
- If $N \lesssim L \lesssim N^2$ then the rate still converges to 0 but we incur the cost $(N^2/L)^{2s/(4s+d)}$ in the minimax rate (trade-off between computational efficiency and rate of convergence).
- If $L \lesssim N$ then there is no guarantee that the rate converges to 0.

To summarize, the tests we propose have a computational cost which can be specified by the user (the size of the design between 1 and N^2), there is a trade-off between test power and computational cost. We provide theoretical rates in terms of the sample size N and the design size L, working up to a constant. The rate is minimax optimal in the case where L grows quadratically with N. We quantify exactly how the rate deteriorates from quadratic (minimax) to linear (no guarantee) growth of the design size with respect to the sample size. In our experiments, we use a design size which grows linearly with the sample size in order to compare our tests against other linear-time tests. However, the assumption guaranteeing a rate converging to 0 is not satisfied in this setting.

6 Efficient aggregated kernel tests using incomplete U-statistics

We now introduce our aggregated tests that combine single tests with different bandwidths. Our aggregation scheme is similar to those in Fromont et al. (2013), Albert et al. (2022) and Schrab et al. (2021, 2022), and can yield an adaptive test to the unknown smoothness parameter s of the Sobolev ball $\mathcal{S}_d^s(R)$, with relatively low price. Let Λ be a finite collection of bandwidths, $(w_\lambda)_{\lambda\in\Lambda}$ be associated weights satisfying $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$ and u_α be some correction term defined shortly in Equation (17). Then, using the incomplete U-statistic $\overline{U}_\lambda$, we define our aggregated test Δ_α^Λ as

$$\Delta_\alpha^\Lambda(\mathbb{Z}_N) := \mathbf{1}\Big(\overline{U}_\lambda(\mathbb{Z}_N) > \widehat{q}_{1-u_\alpha w_\lambda}^\lambda \text{ for some } \lambda \in \Lambda\Big).$$

The levels of the single tests are weighted and adjusted with a correction term

$$u_\alpha := \sup_{B_3}\left\{u \in \left(0, \min_{\lambda\in\Lambda} w_\lambda^{-1}\right) : \frac{1}{B_2}\sum_{b=1}^{B_2}\mathbf{1}\left(\max_{\lambda\in\Lambda}\left(\widetilde{U}_\lambda^b - U_\lambda^{\bullet\lceil B_1(1-uw_\lambda)\rceil}\right) > 0\right) \leq \alpha\right\}, \quad (17)$$

where the wild bootstrapped incomplete U-statistics $\widetilde{U}_\lambda^1, \dots, \widetilde{U}_\lambda^{B_2}$ computed as in Equation (12) are used to perform a Monte Carlo approximation of the probability under the null, and where the supremum is estimated using B_3 steps of bisection method. Proposition 1, along with the reasoning of Schrab et al. (2021, Proposition 8), ensures that Δ_α^Λ has non-asymptotic level α for the two-sample and independence cases, and asymptotic level α for the goodness-of-fit case. We refer to the three aggregated test constructed using incomplete U-statistics as MMDAggInc, HSICAggInc and KSDAggInc. The computational complexity of those tests is $\mathcal{O}\left(|\Lambda|(B_1 + B_2)L\right)$, which means that if $L \asymp N$ as in Equation (11), the tests run efficiently in linear time in the sample size.

We formally record error guarantees of Δ_α^Λ and derive uniform separation rates over Sobolev balls.

Theorem 2. *(i) Suppose that assumptions $\boldsymbol{\mathcal{A}_2}\big(\|p\|_\infty, \|q\|_\infty, \alpha, \beta, B_1, B_2, B_3\big)$ in Appendix A hold and let $\sigma_{2,\lambda}^2 := \mathbb{E}\big[h_\lambda(Z, Z')^2\big]$. Consider a collection Λ such that $\lambda_1 \cdots \lambda_d \lesssim 1$ for all $\lambda \in \Lambda$. For sample size N and design size L, if there exists some $C > 0$ such that*

$$\|p - q\|_2^2 \;\geq\; \min_{\lambda \in \Lambda} \left(\|(p-q) - T_\lambda(p-q)\|_2^2 + C \frac{N}{L} \frac{\ln(1/(\alpha w_\lambda))}{\beta} \sigma_{2,\lambda} \right),$$

then $\mathbb{P}_{\mathcal{H}_1}\big(\Delta_\alpha^\Lambda(\mathbb{Z}_N) = 0\big) \leq \beta$ (type II error), where $\sigma_{2,\lambda} \lesssim 1/\sqrt{\lambda_1 \cdots \lambda_d}$ for MMD and HSIC.

(ii) Consider the collections of bandwidths and weights (independent of R and s)

$$\Lambda := \left\{ \left(2^{-\ell}, \dots, 2^{-\ell}\right) \in (0, \infty)^d : \ell \in \left\{ 1, \dots, \left\lceil \frac{2}{d} \log_2\left(\frac{L/N}{\ln(\ln(L/N))} \right) \right\rceil \right\} \right\}, \quad w_\lambda := \frac{6}{\pi^2 \ell^2}.$$

For two-sample and independence problems, the uniform separation rate of Δ_α^Λ over the Sobolev balls $\left\{\mathcal{S}_d^s(R) : R > 0, s > 0\right\}$ is (up to a constant)

$$\left(\frac{\ln(\ln(L/N))}{L/N} \right)^{2s/(4s+d)}.$$

The extension from Theorem 1 to Theorem 2 has been proved for complete U-statistics in the two-sample (Fromont et al., 2013; Schrab et al., 2021), independence (Albert et al., 2022) and goodness-of-fit (Schrab et al., 2022) testing frameworks. The proof of Theorem 2 follows with the same reasoning by simply replacing N with L/N as we work with incomplete U-statistics; this 'replacement' is theoretically justified by Theorem 1. From Theorem 2, the aggregated test Δ_α^Λ is *adaptive* over Sobolev balls $\left\{\mathcal{S}_d^s(R) : R > 0, s > 0\right\}$: the test Δ_α^Λ does not depend on the unknown smoothness parameter s (unlike $\Delta_\alpha^{\lambda^*}$ in Theorem 1) and achieves the minimax rate up to an iterated logarithmic factor and up to the cost incurred for efficiency of the test (i.e. L/N instead of N).

7 Minimax optimal permuted quadratic-time aggregated independence test

Considering Theorem 2 with our incomplete U-statistic with full design $\mathcal{D} = \mathbf{i}_2^N$ for which $L \asymp N^2$, we have proved that the quadratic-time two-sample and independence aggregated tests using a wild bootstrap achieve the rate $(\ln(\ln(N))/N)^{2s/(4s+d)}$ over the Sobolev balls $\left\{\mathcal{S}_d^s(R) : R > 0, s > 0\right\}$. This is the minimax rate (Li and Yuan, 2019; Albert et al., 2022), up to some iterated logarithmic term. For the two-sample problem, Kim et al. (2022) and Schrab et al. (2021) show that this is also true using complete U-statistics with either a wild bootstrap or permutations. Whether the equivalent statement for independence test with permutations holds is unknown; the rate can be proved using theoretical (unknown) quantiles with a Gaussian kernel (Albert et al., 2022), but has not yet been proved using permutations. Kim et al. (2022, Proposition 8.7) consider this problem, again using a Gaussian kernel, but they do not obtain the correct dependence on α (i.e. $\ln(1/\alpha)$ is replaced with $\alpha^{-1/2}$), hence they cannot recover the desired rate. As pointed out by Kim et al. (2022, Section 8): 'It remains an open question as to whether [the power guarantee] continues to hold when $\alpha^{-1/2}$ is replaced by $\ln(1/\alpha)$'. We now prove that we can improve the α-dependence to $\ln(1/\alpha)^{3/2}$ for any bounded kernel of the form presented in Equation (13), and that this allows us to obtain the desired rate over Sobolev balls $\left\{\mathcal{S}_d^s(R) : R > 0, s > d/4\right\}$. The assumption $s > d/4$ imposes a stronger smoothness restriction on $p - q \in \mathcal{S}_d^s(R)$, which is similarly also considered by Li and Yuan (2019).

Theorem 3. *Consider the quadratic-time independence test using the complete U-statistic HSIC estimator with a quantile estimated using permutations as done by Kim et al. (2022, Proposition 8.7), with kernels as in Equation (13) for bounded functions K_i and L_j for $i = 1, \dots, d_x$, $j = 1, \dots, d_y$.*

(i) Consider the assumptions of Theorem 1. For fixed $R > 0$ and $s > d/4$, with the bandwidths $\lambda_i^* := N^{-2/(4s+d)}$ for $i = 1, \dots, d$, the probability of type II error of the test is controlled by β when

$$\|p - q\|_2^2 \;\geq\; \|(p-q) - T_{\lambda^*}(p-q)\|_2^2 + C\frac{1}{N}\frac{\ln(1/\alpha)^{3/2}}{\beta\sqrt{\lambda_1^* \cdots \lambda_d^*}} \quad \textit{for some constant } C > 0.$$

The uniform separation rate over the Sobolev ball $\mathcal{S}_d^s(R)$ is, up to a constant, $(1/N)^{2s/(4s+d)}$.

(ii) Consider the assumptions of Theorem 2, the uniform separation rate over the Sobolev balls $\left\{\mathcal{S}_d^s(R) : R > 0, s > d/4\right\}$ is $\left(\ln(\ln(N))/N\right)^{2s/(4s+d)}$, up to a constant, with the collections

$$\Lambda := \left\{\left(2^{-\ell}, \dots, 2^{-\ell}\right) \in (0,\infty)^d : \ell \in \left\{1, \dots, \left\lceil \frac{2}{d}\log_2\left(\frac{N}{\ln(\ln(N))}\right)\right\rceil\right\}\right\}, \quad w_\lambda := \frac{6}{\pi^2\ell^2}.$$

The proof of Theorem 3, in Appendix F.5, uses the exponential concentration bound of Kim et al. (2022, Theorem 6.3) for permuted complete U-statistics. Another possible approach to obtaining the correct dependency on α is to employ the sample-splitting method proposed by Kim et al. (2022, Section 8.3) in order to transform the independence problem into a two-sample problem. While this indirect approach leads to a logarithmic factor in α, the practical power would be suboptimal due to an inefficient use of the data from sample splitting. Theorem 3 shows that the usual permutation-based HSIC test has the same logarithmic dependence in α as the less practical sample-splitting test.

8 Experiments

For the two-sample problem, we consider testing samples drawn from a uniform density on $[0,1]^d$ against samples drawn from a perturbed uniform density. For the independence problem, the joint density is a perturbed uniform density on $[0,1]^{d_x+d_y}$, the marginals are then simply uniform densities. Those perturbed uniform densities can be shown to lie in Sobolev balls (Li and Yuan, 2019; Albert et al., 2022), to which our tests are adaptive. For the goodness-of-fit problem, we use a Gaussian-Bernoulli Restricted Boltzmann Machine as first considered by Liu et al. (2016) in this testing framework. We use collections of 21 bandwidths for MMD and HSIC and of 25 bandwidth pairs for HSIC; more details on the experiments (*e.g.* model/test parameters) are presented in Appendix C.

We consider our incomplete aggregated tests MMDAggInc, HSICAggInc and KSDAggInc, with parameter $R \in \{1, \dots, N-1\}$ which fixes the deterministic design to consist of the first R sub-diagonals of the $N \times N$ matrix, that is, $\mathcal{D} := \{(i, i+r) : i = 1, \dots, N-r \text{ for } r = 1, \dots, R\}$ with size $|\mathcal{D}| = RN - R(R-1)/2$. We run our incomplete tests with $R \in \{1, 100, 200\}$ and also consider the complete test which uses the full design $\mathcal{D} = \mathbf{i}_2^N$. We compare their performances with current linear-time state-of-the art tests: OST PSI (Kübler et al., 2020) which performs kernel selection using post selection inference, ME, SCF, FSIC and FSSD (Jitkrittum et al., 2016, 2017a,b) which evaluate the witness functions at a finite set of locations chosen to maximize the power, Cauchy RFF (random Fourier feature) and L1 IMQ (Huggins and Mackey, 2018) which are random feature Stein discrepancies, and LSD (Grathwohl et al., 2020) which uses a neural network to learn the Stein discrepancy (see Appendix C for details).

Similar trends are observed across all our experiments in Figure 1, in the three testing frameworks, when varying the sample size, the dimension, and the difficulty of the problem (scale of perturbations or noise level). The linear-time tests AggInc $R = 200$ almost match the power obtained by the quadratic-time tests AggCom in all settings (except in Figure 1(i) where the difference is larger) while being computationally much more efficient as can be seen in Figure 1(d,h,l). The incomplete tests with $R = 100$ has power only slightly below the one using $R = 200$, and runs roughly twice as fast (Figure 1(d,h,l)). In all experiments, those three tests (AggInc $R = 100, 200$ and AggCom) have significantly higher power than the linear-time tests which optimize test locations (ME, SCF, FSIC and FSSD); in the two-sample case the aggregated tests run faster for small sample size but slower for large sample size, in the independence case the aggregated tests run much faster, and in the goodness-of-fit case the tests optimizing test locations run faster. While both types of tests are linear,

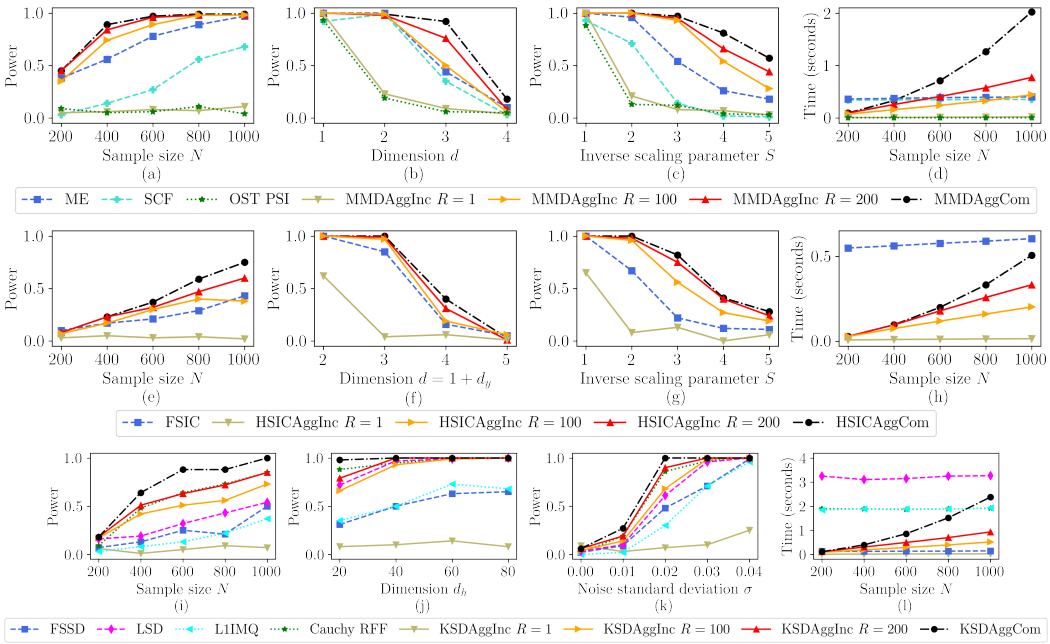

Figure 1: Two-sample *(a–d)* and independence *(e–h)* experiments using perturbed uniform densities. Goodness-of-fit *(i–l)* experiment using a Gaussian-Bernoulli Restricted Boltzmann Machine. The power results are averaged over 100 repetitions and the runtimes over 20 repetitions.

we note that the runtimes of the tests of Jitkrittum et al. (2016, 2017a,b) increase slower with the sample size than our aggregated tests with $R = 100, 200$, but a fixed computational cost is incurred for the optimization step, even for small sample sizes. In the goodness-of-fit framework, L1 IMQ performs similarly to FSSD which is in line with the results presented by Huggins and Mackey (2018, Figure 4d) who consider the same experiment. All other goodness-of-fit tests (except KSDAggInc $R = 1$) achieve much higher test power. Cauchy RFF and KSDAggInc $R = 200$ obtain the same power in almost all the experiments. While KSDAggInc $R = 200$ runs much faster in the experiments presented[6], it seems that the KSDAggInc runtimes increase more steeply with the sample size than the Cauchy RFF and L1 IMQ runtimes (see Appendix D.4 for details). LSD matches the power of KSDAggInc $R = 100$ when varying the noise level in Figure 1(k) (KSDAggInc $R = 200$ has higher power), and matches the power of KSDAggInc $R = 200$ when varying the hidden dimension in Figure 1(j) where $d_x = 100$. When varying the sample size in Figure 1(i), both KSDAggInc tests with $R = 100, 200$ achieve much higher power than LSD. Unsurprisingly, AggInc $R = 1$, which runs much faster than all the aforementioned tests, has low power in every experiment. For the two-sample problem, it obtains slightly higher power than OST PSI which runs even faster. We include additional experiments in Appendix D which verify that all tests have well-calibrated levels, present experiments on the MNIST dataset, and illustrate the benefits of the aggregation procedure.

9 Acknowledgements

Antonin Schrab acknowledges support from the U.K. Research and Innovation (EP/S021566/1). Ilmun Kim acknowledges support from the Yonsei University Research Fund of 2021-22-0332, and from the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2022R1A4A1033384). 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 (EP/R013616/1), and by the French National Agency for Research (ANR-18-CE40-0016-01 & ANR-18-CE23-0015-02). Arthur Gretton acknowledges support from the Gatsby Charitable Foundation.

[6]The runtimes in Figure 1(d,h,l) can also vary due to the different implementations of the respective authors.

References

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

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

Balasubramanian, K., Li, T., and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1).

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

Blom, G. (1976). Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61.

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

Cherfaoui, F., Kadri, H., Anthoine, S., and Ralaivola, L. (2022). A discrete RKHS standpoint for Nyström MMD. *HAL preprint hal-03651849*.

Chwialkowski, K., Sejdinovic, D., and Gretton, A. (2014). A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pages 3608–3616.

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

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

Clémençon, S., Robbiano, S., and Tressou, J. (2013). Maximal deviations of incomplete U-statistics with applications to empirical risk sampling. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 19–27. SIAM.

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

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. In *International Conference on Learning Representations*.

Duembgen, L. (1998). Symmetrization and decoupling of combinatorial random elements. *Statistics & probability letters*, 39(4):355–361.

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

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

Freidling, T., Poignard, B., Climente-González, H., and Yamada, M. (2021). Post-selection inference with HSIC-Lasso. In *International Conference on Machine Learning*, pages 3439–3448. PMLR.

Fromont, M., Laurent, B., Lerasle, M., and Reynaud-Bouret, P. (2012). Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Conference on Learning Theory*, PMLR.

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

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

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. (2020). Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pages 3732–3747. PMLR.

Gretton, A. (2015). A simpler condition for consistency of a kernel independence test. *arXiv preprint arXiv:1501.06103*.

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

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, 22.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 1, pages 585–592.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 1, pages 1205–1213.

Ho, H.-C. and Shieh, G. S. (2006). Two-stage U-statistics for hypothesis testing. *Scandinavian journal of statistics*, 33(4):861–873.

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

Huggins, J. and Mackey, L. (2018). Random feature Stein discrepancies. *Advances in Neural Information Processing Systems*, 31.

Janson, S. (1984). The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(4):495–505.

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

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017a). An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning (ICML)*, pages 1742–1751.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017b). A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271.

Key, O., Fernandez, T., Gretton, A., and Briol, F.-X. (2021). Composite goodness-of-fit tests with kernels. *arXiv preprint arXiv:2111.10275*.

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

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245.

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

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

LeCun, Y., Cortes, C., and Burges, C. (2010). MNIST handwritten digit database. AT&T Labs.

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

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

Leucht, A. and Neumann, M. H. (2013). Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280.

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

Lim, J. N., Yamada, M., Jitkrittum, W., Terada, Y., Matsui, S., and Shimodaira, H. (2020). More powerful selective kernel tests for feature selection. In *International Conference on Artificial Intelligence and Statistics*, pages 820–830. PMLR.

Lim, J. N., Yamada, M., Schölkopf, B., and Jitkrittum, W. (2019). Kernel Stein tests for multiple model comparison. In *Advances in Neural Information Processing Systems*, pages 2240–2250.

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

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

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

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

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

Schrab, A., Guedj, B., and Gretton, A. (2022). KSD aggregated goodness-of-fit test. *arXiv preprint arXiv:2202.00824*.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2021). MMD aggregated two-sample test. *arXiv preprint arXiv:2110.15073*.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235.

Song, L., Smola, A. J., Gretton, A., Bedo, J., and Borgwardt, K. M. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434.

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

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

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2018). Post selection inference with kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 152–160. PMLR.

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

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in neural information processing systems*, 26.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2018). Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130.

Zhao, J. and Meng, D. (2015). FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See framed bullet points at the end of Section 5 (limitation for the case $L < N$).
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See main text for Theorem 2 and appendices for all other proofs.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] URL in Section 1: `https://github.com/antoninschrab/agginc-paper`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For the time plots in Figure 1 we report the error bars which represent the standard deviation (they are really small and cannot always been seen). For the power plots in Figure 1, since the test outputs are binary (0 or 1), there is no need to include error bars since these are deterministic given the average which is plotted.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix C.
   (b) Did you mention the license of the assets? [Yes] See Appendix C.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] URL in Section 1: `https://github.com/antoninschrab/agginc-paper`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Assumptions

The assumptions $\boldsymbol{\mathcal{A}_1}\big(\|p\|_\infty, \|q\|_\infty, \alpha, \beta, B_1\big)$ for Theorem 1 are:

- $\max\left(\|p\|_\infty, \|p\|_\infty\right) \leq M$ for some $M > 0$,
- $\alpha \in (0, e^{-1})$,
- $\beta \in (0, 1)$,
- $B_1 \geq \frac{2}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$.

The assumptions $\boldsymbol{\mathcal{A}_2}\big(\|p\|_\infty, \|q\|_\infty, \alpha, \beta, B_1, B_2, B_3\big)$ for Theorem 2 are:

- $\max\left(\|p\|_\infty, \|p\|_\infty\right) \leq M$ for some $M > 0$,
- $\alpha \in (0, e^{-1})$,
- $\beta \in (0, 1)$,
- $B_1 \geq \frac{2}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$,
- $B_2 \geq \frac{8}{\alpha^2}\ln\big(\frac{2}{\beta}\big)$,
- $B_3 \geq \log_2\big(\frac{4}{\alpha}\min_{\lambda\in\Lambda} w_\lambda^{-1}\big)$.

B Background details on MMD, HSIC and KSD

In this section, we present more details than those presented in Section 2 on the Maximum Mean Discrepancy, on the Hilbert Schmidt Independence Criterion, and on the Kernel Stein Discrepancy.

Maximum Mean Discrepancy. Gretton et al. (2012a) introduce the *Maximum Mean Discrepancy* (MMD) which is a measure between probability densities p and q on $\mathbb{R}^d$. It is defined as the integral probability metric (IPM; Müller, 1997) over a reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS; Aronszajn, 1950) with associated kernel k. Gretton et al. (2012a, Lemma 4) show that the MMD is equal to 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 $u \in \mathbb{R}^d$. The square of the MMD is equal to

$$\begin{aligned}\mathrm{MMD}_k^2(p,q) &\coloneqq \Big(\sup_{f\in\mathcal{H}_k\,:\,\|f\|_{\mathcal{H}_k}\leq 1} \big|\mathbb{E}_p[f(X)] - \mathbb{E}_q[f(Y)]\big|\Big)^2 \\ &= \|\mu_p - \mu_q\|_{\mathcal{H}_k}^2 \\ &= \mathbb{E}_{p,p}\big[k\big(X,X'\big)\big] - 2\,\mathbb{E}_{p,q}\big[k\big(X,Y\big)\big] + \mathbb{E}_{q,q}\big[k\big(Y,Y'\big)\big]\end{aligned}$$

where X and X' (respectively Y and Y') are independent. Using a characteristic kernel (Fukumizu et al., 2008; Sriperumbudur et al., 2011) guarantees that $\mathrm{MMD}_k^2(p,q) = 0$ if and only if $p = q$, a crucial property for using the MMD to construct a two-sample test. With i.i.d. samples $\mathbb{X}_m \coloneqq (X_i)_{1\leq i\leq m}$ from p and i.i.d. samples $\mathbb{Y}_n = (Y_j)_{1\leq j\leq n}$ from q, Gretton et al. (2012a, Lemma 6) propose to use the unbiased quadratic-time MMD estimator $\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n)$ defined as

$$\begin{aligned}&\frac{1}{m(m-1)}\sum_{(i,i')\in\mathbf{i}_2^m} k(X_i,X_{i'}) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(X_i,Y_j) + \frac{1}{n(n-1)}\sum_{(j,j')\in\mathbf{i}_2^n} k(Y_j,Y_{j'}) \\ &= \frac{\mathbf{1}^\top\tilde{\mathbf{K}}_{\mathrm{XX}}\mathbf{1}}{m(m-1)} - 2\frac{\mathbf{1}^\top\mathbf{K}_{\mathrm{XY}}\mathbf{1}}{mn} + \frac{\mathbf{1}^\top\tilde{\mathbf{K}}_{\mathrm{YY}}\mathbf{1}}{n(n-1)}\end{aligned}$$

where $\tilde{\mathbf{K}}_{\mathrm{XX}}$ and $\tilde{\mathbf{K}}_{\mathrm{YY}}$ are the kernel matrices $\mathbf{K}_{\mathrm{XX}} \coloneqq \big(k(X_i,X_j)\big)_{1\leq i,j\leq m}$ and $\mathbf{K}_{\mathrm{YY}} \coloneqq \big(k(Y_i,Y_j)\big)_{1\leq i,j\leq n}$ with diagonal entries set to 0, where $\mathbf{K}_{\mathrm{XY}} \coloneqq \big(k(X_i,Y_j)\big)_{1\leq i\leq m, 1\leq j\leq n}$, and where $\mathbf{1}$ is a one-dimensional vector with all entries equal to 1 of variable length determined by the context[7]. As noted by Kim et al. (2022), this MMD estimator can be rewritten as a two-sample

[7] We use this convention for the notation $\mathbf{1}$ in this whole section.

U-statistic (both of second order) (Hoeffding, 1992)

$$\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n) = \frac{1}{\left|\mathbf{i}_2^m\right|\left|\mathbf{i}_2^n\right|} \sum_{(i,i')\in\mathbf{i}_2^m} \sum_{(j,j')\in\mathbf{i}_2^n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_j, Y_{j'})$$

where $\mathbf{i}_a^b$ denotes the set of all a-tuples drawn without replacement from $\{1, \dots, b\}$ so that $\left|\mathbf{i}_a^b\right| = b \cdots (b-a+1)$, for example $\left|\mathbf{i}_2^m\right| = m(m-1)$, and where, for $x_1, x_2, y_1, y_2 \in \mathbb{R}^d$, we let

$$h_k^{\mathrm{MMD}}(x_1, x_2; y_1, y_2) \coloneqq k(x_1, x_2) - k(x_1, y_2) - k(x_2, y_1) + k(y_1, y_2).$$

This kernel can easily be symmetrized (Kim et al., 2022) using a symmetrization trick (Duembgen, 1998), this corresponds to working with

$$\bar{h}_k^{\mathrm{MMD}}(x_1, x_2; y_1, y_2) \coloneqq \frac{1}{2!2!} \sum_{(i_1,i_2)\in\mathbf{i}_2^2} \sum_{(j_1,j_2)\in\mathbf{i}_2^2} h_k^{\mathrm{MMD}}(x_1, x_2; y_1, y_2)$$

and the MMD expression as a U-statistic still holds when replacing h_k^{MMD} with its symmetrized variant $\bar{h}_k^{\mathrm{MMD}}$.

Hilbert Schmidt Independence Criterion. For a joint probability density p_{xy} on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ with marginals p_x on $\mathbb{R}^{d_x}$ and p_y on $\mathbb{R}^{d_y}$, Gretton et al. (2005) introduce the *Hilbert Schmidt Independence Criterion* (HSIC) which is defined as

$$\begin{aligned}\mathrm{HSIC}_{k,\ell}(p_{xy}) &\coloneqq \mathrm{MMD}_\kappa^2(p_{xy}, p_x p_y) \\ &= \mathbb{E}_{p_{xy},p_{xy}}\Big[k(X,X')\ell(Y,Y')\Big] - 2\,\mathbb{E}_{p_{xy}}\Big[\mathbb{E}_{p_x}[k(X,X')]\mathbb{E}_{p_y}[\ell(Y,Y')]\Big] \\ &\qquad + \mathbb{E}_{p_x,p_x}\Big[k(X,X')\Big]\mathbb{E}_{p_y,p_y}\Big[k(Y,Y')\Big].\end{aligned}$$

with kernels k on $\mathbb{R}^{d_x}$ and ℓ on $\mathbb{R}^{d_y}$ giving the product kernel $\kappa\big((x,y),(x',y')\big) \coloneqq k(x,x')\ell(y,y')$ on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. With i.i.d. pairs of samples $\mathbb{Z}_N \coloneqq \big(Z_i\big)_{1\le i\le N} = \big((X_i, Y_i)\big)_{1\le i\le N}$ drawn from p_{xy}, a natural unbiased HSIC estimator (Gretton et al., 2008; Song et al., 2012) is then

$$\begin{aligned}\widehat{\mathrm{HSIC}}_{k,\ell}(\mathbb{Z}_N) &\coloneqq \frac{1}{\left|\mathbf{i}_2^N\right|} \sum_{(i,j)\in\mathbf{i}_2^N} k(X_i,X_j)\ell(Y_i,Y_j) - \frac{2}{\left|\mathbf{i}_3^N\right|} \sum_{(i,j,r)\in\mathbf{i}_3^N} k(X_i,X_j)\ell(Y_i,Y_r) \\ &\qquad + \frac{1}{\left|\mathbf{i}_4^N\right|} \sum_{(i,j,r,s)\in\mathbf{i}_4^N} k(X_i,X_j)\ell(Y_r,Y_s) \\ &= \frac{1}{\left|\mathbf{i}_4^N\right|} \sum_{(i,j,r,s)\in\mathbf{i}_4^N} h_{k,\ell}^{\mathrm{HSIC}}(Z_i, Z_j, Z_r, Z_s)\end{aligned}$$

which is a fourth-order one-sample U-statistic. For $z_a = (x_a, y_a) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $a = 1, \dots, 4$, we let

$$h_{k,\ell}^{\mathrm{HSIC}}(z_1, z_2, z_3, z_4) \coloneqq \frac{1}{4} h_k^{\mathrm{MMD}}(x_1, x_2; x_3, x_4) h_\ell^{\mathrm{MMD}}(y_1, y_2; y_3, y_4).$$

We stress the fact that this HSIC estimator can actually be computed in quadratic time as shown by Song et al. (2012, Equation 5) who provide the following closed-form expression

$$\widehat{\mathrm{HSIC}}_{k,\ell}(\mathbb{Z}_N) = \frac{1}{N(N-3)}\left(\mathrm{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) + \frac{\mathbf{1}^\top\tilde{\mathbf{K}}\mathbf{1}\mathbf{1}^\top\tilde{\mathbf{L}}\mathbf{1}}{(N-1)(N-2)} - \frac{2}{N-2}\mathbf{1}^\top\tilde{\mathbf{K}}\tilde{\mathbf{L}}\mathbf{1}\right)$$

where $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ are the kernel matrices $\mathbf{K} \coloneqq \big(k(X_i, X_j)\big)_{1\le i,j\le N}$ and $\mathbf{L} \coloneqq \big(\ell(Y_i, Y_j)\big)_{1\le i,j\le N}$ with diagonal entries set to 0. Again, this kernel can be symmetrized (Song et al., 2012; Kim et al., 2022) using a symmetrization trick (Duembgen, 1998), and the HSIC expression as a U-statistic still holds when replacing h_k^{HSIC} with its symmetrized variant

$$\bar{h}_k^{\mathrm{HSIC}}(z_1, z_2, z_3, z_4) \coloneqq \frac{1}{4!} \sum_{(i_1,i_2,i_3,i_4)\in\mathbf{i}_4^4} h_k^{\mathrm{HSIC}}(z_{i_1}, z_{i_2}, z_{i_3}, z_{i_4}).$$

Kernel Stein Discrepancy. For probability densities p and q on $\mathbb{R}^d$, Chwialkowski et al. (2016) and Liu et al. (2016) introduce the *Kernel Stein Discrepancy* (KSD) defined as

$$\begin{aligned}\mathrm{KSD}^2_{p,k}(q) &\coloneqq \mathrm{MMD}^2_{h^{\mathrm{KSD}}_{k,p}}(q,p)\\ &= \mathbb{E}_{q,q}\left[h^{\mathrm{KSD}}_{k,p}(Z,Z')\right] - 2\,\mathbb{E}_{q,p}\left[h^{\mathrm{KSD}}_{k,p}(Z,X)\right] + \mathbb{E}_{p,p}\left[h^{\mathrm{KSD}}_{k,p}(X,X')\right]\\ &= \mathbb{E}_{q,q}\left[h^{\mathrm{KSD}}_{k,p}(Z,Z')\right]\end{aligned}$$

where the *Stein kernel* $h_{p,k}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{aligned}h^{\mathrm{KSD}}_{k,p}(x,y) \coloneqq\ & \left(\nabla \log p(x)^\top \nabla \log p(y)\right) k(x,y) + \nabla \log p(y)^\top \nabla_x k(x,y)\\ & + \nabla \log p(x)^\top \nabla_y k(x,y) + \sum_{i=1}^d \frac{\partial}{\partial x_i \partial y_i} k(x,y).\end{aligned}$$

The Stein kernel satisfies the Stein identity $\mathbb{E}_p[h^{\mathrm{KSD}}_{k,p}(X,\cdot)] = 0$. The KSD is particularly useful for the goodness-of-fit setting with a model density p and i.i.d. samples $\mathbb{Z}_N \coloneqq (Z_i)_{1\leq i\leq N}$ drawn from a density q because it admits an estimator which does not require samples from the model p. The quadratic-time KSD estimator can be computed as the second-order one-sample U-statistic

$$\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) \coloneqq \frac{1}{\left|\mathbf{i}^N_2\right|} \sum_{(i,j)\in\mathbf{i}^N_2} h^{\mathrm{KSD}}_{k,p}(Z_i,Z_j) = \frac{\mathbf{1}^\top \tilde{\mathbf{H}} \mathbf{1}}{N(N-1)}$$

where $\tilde{\mathbf{H}}$ is the kernel matrix $\mathbf{H} \coloneqq \left(h^{\mathrm{KSD}}_{k,p}(Z_i,Z_j)\right)_{1\leq i,j\leq N}$ with diagonal entries set to 0. The Stein kernel h^{KSD}_k is already symmetric, we can write $\bar{h}^{\mathrm{KSD}}_k(x,y) \coloneqq h^{\mathrm{KSD}}_k(x,y)$ for all $x,y \in \mathbb{R}^d$ for consistency of notation.

Quantile estimation. There exist many approaches to estimating the quantiles of the test statistics under the null hypothesis in the three frameworks: using the quantile of a known distribution-free asymptotic null distribution (Gretton et al., 2008, 2012a), sampling from an asymptotic null distribution with eigenspectrum approximation (Gretton et al., 2009), using permutations (Gretton et al., 2008; Albert et al., 2022; Kim et al., 2022; Schrab et al., 2021), using a wild bootstrap (Fromont et al., 2012; Chwialkowski et al., 2014, 2016; Schrab et al., 2021, 2022), using parametric bootstrap (Key et al., 2021; Schrab et al., 2022), using other bootstrap methods (Liu et al., 2016), to name but a few. Permutation-based tests have been shown to correctly control the non-asymptotic level for the two-sample (Schrab et al., 2021; Kim et al., 2022) and independence (Albert et al., 2022; Kim et al., 2022) problems. For the two-sample test, using a wild bootstrap also guarantees well-calibrated non-asymptotic level (Fromont et al., 2012; Schrab et al., 2021). For the goodness-of-fit setting, while a wild bootstrap guarantees only control of the asymptotic level (Chwialkowski et al., 2016), using a parametric bootstrap results in well-calibrated non-asymptotic level (Schrab et al., 2022). In this work, we focus on the wild bootstrap approach, though we point out that our results also hold using a parametric bootstrap for the goodness-of-fit setting as done by Schrab et al. (2022).

C Detailed experimental protocol

In this section, we present details on the experiments and on the tests considered. We also show empirically that all the tests have well-calibrated levels.

Implementation and computational resources. All experiments have been run on an AMD Ryzen Threadripper 3960X 24 Cores 128Gb RAM CPU at 3.8GHz, except the LSD test (Grathwohl et al., 2020) for which a neural network has been trained using a NVIDIA RTX A5000 24Gb Graphics Card. The overall runtime of all the experiments is of the order of a couple of hours (significant speedup can be obtained by using parallel computing). For the ME, SCF, FSIC and FSSD tests of Jitkrittum et al. (2016, 2017a,b), for the LSD test of Grathwohl et al. (2020), for the L1 IMQ & Cauchy RFF test of Huggins and Mackey (2018), and for the OST PSI test of Kübler et al. (2020), we use the implementations of the respective authors, all under the MIT license. The implementation of our computationally efficient aggregated tests, as well as the code for reproducibility of the experiments, are available here under the MIT license.

Kernels. For the two-sample and independence experiments, we use the Gaussian kernel[8] with equal bandwidths $\lambda_1 = \cdots = \lambda_d = \tilde{\lambda}$, which is defined as

$$k_\lambda(x, y) \coloneqq \exp\left(-\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{\lambda_i^2}\right) = \exp\left(-\frac{\|x - y\|_2^2}{\tilde{\lambda}^2}\right),$$

and similarly for the kernel ℓ_μ. As shown by Gorham and Mackey (2017), a more appropriate kernel for goodness-of-fit testing is the IMQ (inverse multiquadric) kernel

$$k_\lambda(x, y) \coloneqq \left(1 + \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{\lambda_i^2}\right)^{-\beta_k} = \left(1 + \frac{\|x - y\|_2^2}{\tilde{\lambda}^2}\right)^{-\beta_k} \propto \left(\tilde{\lambda}^2 + \|x - y\|_2^2\right)^{-\beta_k}$$

for some $\beta_k \in (0, 1)$. In our goodness-of-fit experiments, we fix the parameter $\beta_k = 0.5$.

Two-sample and independence experiments. In our experiments, we consider perturbed uniform densities, those can be shown to lie in Sobolev balls and are used to derive the minimax rates over Sobolev balls for the two-sample and independence problems (Albert et al., 2022; Li and Yuan, 2019). For the two-sample problem, we consider testing samples drawn from a uniform density against samples drawn from a perturbed uniform density, as considered by Schrab et al. (2021, see Equation 17 for formal definition and Figure 2 for illustrations). We scale the perturbations so that the perturbed density takes value in the whole interval $[0, 2]$, we then consider some inverse scaling parameter $S \geq 1$ such that it takes value in the interval $[1 - 1/S, 1 + 1/S]$. Intuitively, as S increases, the perturbation is shrunk. In Figure 1(a,d), we consider 2 perturbations with inverse scaling parameter $S = 2$ in dimension $d = 1$ and vary the sample size $N \in \{200, 400, 600, 800, 1000\}$. In Figure 1(b), we vary the dimension $d \in \{1, 2, 3, 4\}$ for 1 perturbation with $S = 1$ and $N = 1000$. In Figure 1(c), we use 1 perturbation with $d = 1$ and $N = 1000$, we vary the inverse scaling parameter $S \in \{1, 2, 3, 4, 5\}$. For the independence problem, we draw samples from the joint perturbed uniform density in dimension $d_x + d_y$, the marginals are simply uniform densities in dimensions d_x and d_y, respectively. We fix $d_x = 1$ and vary d_y exactly as in the two-sample setting. The parameters for the independence experiments in Figure 1(e–h) are the same as those of the two-sample experiments in Figure 1(a–d) detailed above (with the only difference that for Figure 1(f) we consider $d_y \in \{1, 2, 3\}$).

Goodness-of-fit experiments. In Figure 1(i–l), we use a Gaussian-Bernoulli Restricted Boltzmann Machine (GBRBM) with the same setting considered by Liu et al. (2016), Grathwohl et al. (2020) and Schrab et al. (2022). This is a hidden variable model with a continuous observable variable in $\mathbb{R}^{d_x}$ and a hidden binary variable in $\{-1, 1\}^{d_h}$, the joint density is intractable but the score function admits a closed form. The GBRBM has parameters $b \in \mathbb{R}^{d_x}$ and $c \in \mathbb{R}^{d_h}$, which are drawn from Gaussian standard distributions, and a parameter $B \in \mathbb{R}^{d_x \times d_h}$. For the model p, the elements of B are sampled uniformly from $\{-1, 1\}$ (i.i.d. Rademacher variables). The samples come from a GBRBM q with the same parameters as the model p but where some Gaussian noise $\mathcal{N}(0, \sigma)$ is injected into the elements of B. In Figure 1(i,l), we consider dimensions $d_x = 50$ and $d_h = 40$ with noise standard deviation $\sigma = 0.02$ and we vary the sample size $N \in \{200, 400, 600, 800, 1000\}$. In Figure 1(j), we fix $d_x = 100$, $N = 1000$, $\sigma = 0.03$ and we vary the hidden dimension $d_h \in \{20, 40, 60, 80\}$. For fixed observed dimension d_x, as the hidden dimension d_h increases the size of $B \in \mathbb{R}^{d_x \times d_h}$ becomes larger, so there is more evidence of the noise being injected, which makes the problem easier. Hence, the test power increases as d_h increases for fixed d_x. In Figure 1(k), we consider dimensions $d_x = 50$ and $d_h = 40$ with sample size $N = 1000$, we vary the noise standard deviations $\sigma \in \{0, 0.01, 0.02, 0.03, 0.04\}$.

AggInc tests. For MMDAggInc and KSDAggInc, we use the collection of 21 bandwidths

$$\Lambda \coloneqq \left\{2^i \lambda_{\text{med}} \mathbf{1}_d : i \in \{-10, \ldots, 10\}\right\} \quad \text{where} \quad \lambda_{\text{med}} \coloneqq \text{median}\left\{\|z_i - z_j\|_2 : (i, j) \in \mathbf{i}_2^N\right\}.$$

where $\mathbf{1}_d$ is a d-dimensional vector with all entries equal to 1. For HSICAggInc, we work with the collection of pairs of 25 bandwidths

$$\Lambda \coloneqq \left\{\left(2^i \lambda_{\text{med}} \mathbf{1}_{d_x}, 2^j \mu_{\text{med}} \mathbf{1}_{d_y}\right) : i, j \in \{-2, -1, 0, 1, 2\}\right\}$$

[8] In practice, we do not need to normalize the kernels to integrate to 1 since our tests are invariant to multiplying the kernel by a scalar.

for the kernels k_λ and ℓ_μ defined in Equation (13), where

$$\lambda_{\text{med}} := \text{median}\Big\{\|x_i - x_j\|_2 : (i,j) \in \mathbf{i}_2^N\Big\} \quad \text{and} \quad \mu_{\text{med}} := \text{median}\Big\{\|y_i - y_j\|_2 : (i,j) \in \mathbf{i}_2^N\Big\}.$$

In practice, depending on our computational budget, we can also consider multiple kernels, each with various bandwidths. All aggregated tests are run with uniform weights defined as $w_\lambda := 1/|\Lambda|$ for all $\lambda \in \Lambda$. The design choice consists of R sub-diagonals of the kernel matrix for $R \in \{1, 100, 200\}$, it is formally defined in Section 8. We also consider the quadratic-time case where the full design is considered (*i.e.* case $R = N - 1$), we refer to these tests using complete U-statistics as AggCom for consistency. We note that MMDAggCom, HSICAggCom and KSDAggCom simply correspond to the quadratic-time MMDAgg, HSICAgg and KSDAgg tests proposed by Schrab et al. (2021), Albert et al. (2022) and Schrab et al. (2022), respectively, with the only difference being their implementation: Agg tests run slightly faster than AggCom tests since they can exploit the fact that the whole kernel matrix needs to be computed. We use $B_1 = 500$ and $B_2 = 500$ wild bootstrapped statistics to estimate the quantiles and the probability under the null for the correction in Equation (17), respectively. For that correction term, we use $B_3 = 50$ steps of bisection method to approximate the supremum.

ME, SCF, FSIC and FSSD tests. Jitkrittum et al. (2016) uses the two-sample tests ME and SCF proposed by Chwialkowski et al. (2015) with features which are chosen to maximise a lower bound on the test power. The ME test is based on analytic Mean Embeddings while the SCF test uses the difference in Smooth Characteristic Functions. For the independence problem, Jitkrittum et al. (2017a) constructs a FSIC test which uses their proposed normalised Finite Set Independence Criterion. Jitkrittum et al. (2017b) proposes a goodness-of-fit test based on the Finite Set Stein Discrepancy (FSSD). All those tests utilise test statistics which evaluate the witness function of either MMD, HSIC, or KSD, at some test locations (i.e. features) chosen on held-out data to maximise test power. For the two-sample SCF test, the test locations are in the frequency domain rather than in the spatial domain. All tests are used with 10 test locations which are chosen on half of the data, as done in the experiments of Jitkrittum et al. (2016, 2017a,b). The ME and SCF tests use the quantiles of the known chi-square asymptotic null distributions. The FSIC test uses 500 permutations to simulate the null hypothesis and compute the test threshold to ensure well-calibrated non-asymptotic level. The FSSD test simulates 2000 samples from the asymptotic null distribution (weighted sum of chi-squares) with the eigenvalues being computed from the covariance matrix with respect to the observed sample. For the two-sample and independence tests, the bandwidths of the Gaussian kernels are selected during the optimization procedure. For the goodness-of-fit test, the bandwidth of the IMQ (inverse multiquadric) kernel is set to one as done by Jitkrittum et al. (2017b), following the recommendation of Gorham and Mackey (2017).

LSD test. The Kernelised Stein Discrepancy (KSD) is a Stein Discrepancy (Gorham and Mackey, 2017) where the class of functions is taken to be the unit ball of a reproducing kernel Hilbert space (RKHS). Grathwohl et al. (2020) propose to instead consider some more expressive class of functions consisting of neural networks, resulting in the Learned Stein Discrepancy (LSD). For goodness-of-fit testing, they propose to split the data into training (80%), validation (10%) and testing (10%) sets. They construct a test statistic which is asymptotically normal under both $\mathcal{H}_0$ and $\mathcal{H}_1$. Using the training set, they train the parametrised neural network to maximise the test power by optimizing a proxy for it which is derived following the reasonings of Gretton et al. (2012b), Sutherland et al. (2017) and Jitkrittum et al. (2017b). They perform model selection on the validation set. Finally, they run the test on the testing set using the quantile of the asymptotic normal distribution under the null. As in the experiments of Grathwohl et al. (2020), a 2-layer MLP with 300 units per layer and with Swish nonlinearity (Ramachandran et al., 2017) is used. Their model is trained using the Adam optimizer (Kingma and Ba, 2014) for 1000 iterations, with dropout (Srivastava et al., 2014), with weight decay of strength 0.0005, with learning rate 10^{-3}, and with L^2 regularising strength 0.5.

Cauchy RFF and L1 IMQ tests. Huggins and Mackey (2018) introduce random feature Stein discrepancies (RΦSDs) which are computable in linear time. The FSSD of Jitkrittum et al. (2017b) corresponds to a specific RΦSD. Another special case of their general RΦSDs is the random Fourier feature (RFF; Rahimi and Recht, 2007) approximation of KSD. They consider in their experiments both Gaussian and Cauchy RFF tests, they observe that Cauchy RFF significantly outperforms its Gaussian counterpart (Huggins and Mackey, 2018, Figure 4). Using the inverse multiquadric kernel (IMQ), for which Gorham and Mackey (2017) showed that KSD dominates weak convergence when $\beta \in (-1, 0)$, Huggins and Mackey (2018, Example 3.4) derive a L^r IMQ RΦSD, with some simple setting when $r = 1$. They show in their experiments that L1 IMQ has superior performance

compared to all other tests considered for experiments comparing Gaussian and Laplace distributions, as well as Gaussian and multivariate t distributions. We use the parameters recommended by the authors when running Cauchy RFF and L1 IMQ, except for the number of samples drawn from the unnormalized density to estimate the covariance matrix to simulate the null hypothesis. As explained in Appendix D.4, we tune that number in order for their tests to be more computationally efficient while retaining their high test power.

OST PSI test. Kübler et al. (2020) construct an MMD adaptive two-sample test which uses the post-selection inference framework (PSI; Fithian et al., 2014; Lee et al., 2016) (with uncountable candidate sets) to use the same data to both perform kernel selection and run the test while still guaranteeing control of the probability of type I error. Their one-sided test (OST) run in linear time and does not rely on data splitting. For kernel selection, they use a proxy for asymptotic power as a criterion. We use their implementation with the same collection of bandwidths as for our MMDAggInc as specified above.

D Additional experiments

In this section, we present additional experiments. We show that the tests we consider are all well-calibrated. We present experiments highlighting the strengths of the aggregation procedure. We also consider more challenging experiments on the high-dimensional MNIST dataset. Finally, we discuss the choice of parameters for the L1 IMQ and Cauchy RFF tests.

D.1 Well-calibrated levels

All tests are run with level $\alpha = 0.05$, it is verified in Tables 1 to 6 that all tests have well-calibrated levels for the three testing frameworks, when varying either the sample size or the dimension. The levels plotted are averages obtained across 200 repetitions, this explains the small fluctuations observed from the desired test level $\alpha = 0.05$. The settings of those six experiments correspond to the settings of the experiments presented in Figure 1(a,b,e,f,i,j) detailed above, with the difference that we are working under the null hypothesis (i.e. perturbed uniform densities are replaced with uniform densities, and the noise standard deviation for the GBRBB is set to $\sigma = 0$).

Table 1: Two-sample level experiment using uniform densities varying the sample size.

Sample size	ME	SCF	OST PSI	MMDAggInc $R = 1$	MMDAggInc $R = 100$	MMDAggInc $R = 200$	MMDAggCom
200	0.055	0.005	0.045	0.04	0.05	0.055	0.055
400	0.08	0.01	0.04	0.035	0.06	0.03	0.03
600	0.08	0.005	0.105	0.085	0.04	0.04	0.07
800	0.05	0.005	0.055	0.075	0.03	0.035	0.055
1000	0.075	0.005	0.045	0.045	0.015	0.02	0.05

Table 2: Two-sample level experiment using uniform densities varying the dimension.

Dimension	ME	SCF	OST PSI	MMDAggInc $R = 1$	MMDAggInc $R = 100$	MMDAggInc $R = 200$	MMDAggCom
1	0.045	0	0.035	0.02	0.045	0.04	0.045
2	0.045	0.035	0.085	0.1	0.05	0.04	0.035
3	0.04	0.05	0.04	0.04	0.05	0.06	0.025
4	0.045	0.05	0.03	0.055	0.045	0.045	0.03

Table 3: Independence level experiment using uniform densities varying the sample size.

Sample size	FSIC	HSICAggInc $R = 1$	HSICAggInc $R = 100$	HSICAggInc $R = 200$	HSICAggCom
200	0.04	0.055	0.035	0.035	0.035
400	0.045	0.05	0.04	0.05	0.05
600	0.05	0.035	0.05	0.06	0.05
800	0.03	0.07	0.02	0.035	0.04
1000	0.07	0.02	0.085	0.035	0.04

Table 4: Independence level experiment using uniform densities varying the dimension.

Dimension	FSIC	HSICAggInc $R = 1$	HSICAggInc $R = 100$	HSICAggInc $R = 200$	HSICAggCom
2	0.035	0.065	0.08	0.055	0.07
3	0.065	0.055	0.035	0.02	0.025
4	0.04	0.035	0.045	0.055	0.055

Table 5: Goodness-of-fit level experiment using a Gaussian-Bernoulli Restricted Boltzmann Machine varying the sample size.

Sample size	FSSD	LSD	KSDAggInc $R = 1$	KSDAggInc $R = 100$	KSDAggInc $R = 200$	KSDAggCom
200	0.02	0.07	0.05	0.045	0.06	0.06
400	0.03	0.04	0.06	0.04	0.065	0.055
600	0.04	0.075	0.03	0.03	0.04	0.07
800	0.03	0.06	0.055	0.06	0.045	0.07
1000	0.025	0.05	0.045	0.035	0.045	0.065

Table 6: Goodness-of-fit level experiment using a Gaussian-Bernoulli Restricted Boltzmann Machine varying the dimension.

Dimension	FSSD	LSD	KSDAggInc $R = 1$	KSDAggInc $R = 100$	KSDAggInc $R = 200$	KSDAggCom
20	0.02	0.055	0.045	0.06	0.065	0.05
40	0.04	0.055	0.07	0.055	0.065	0.07
60	0.04	0.055	0.06	0.04	0.05	0.06
80	0.015	0.04	0.045	0.04	0.035	0.05

D.2 Aggregation experiments

We illustrate in Figure 2 the benefits of the aggregation procedure by starting from a 'collection' consisting of only the median bandwidth and increasing the collection by adding more bandwidths. In all three settings, we observe that the power for the test with the median bandwidth only is low. As we increase the number of bandwidths, the power first increases as the test has access to 'better-suited' bandwidths.

For MMDAggInc and KSDAggInc, once the optimal bandwidth is included in the bandwidth, the power reaches a plateau. We do not pay a price in power for considering more bandwidths (or kernels), and so the user is encouraged to consider many kernels with various bandwidths. For the unscaled Gaussian kernel, we are essentially aggregating over kernel matrices which interpolate between the identity matrix (as the bandwidth goes to 0) and the matrix of ones (as the bandwidth goes to ∞).

HSICAggInc is more challenging, since there are pairs of kernels, hence the total number of bandwidth combinations grows rapidly (*e.g.* b bandwidths for each kernel corresponds to b^2 pairs of kernels). In this case, we observe a significant decay in test power once more than 49 bandwidths are considered.

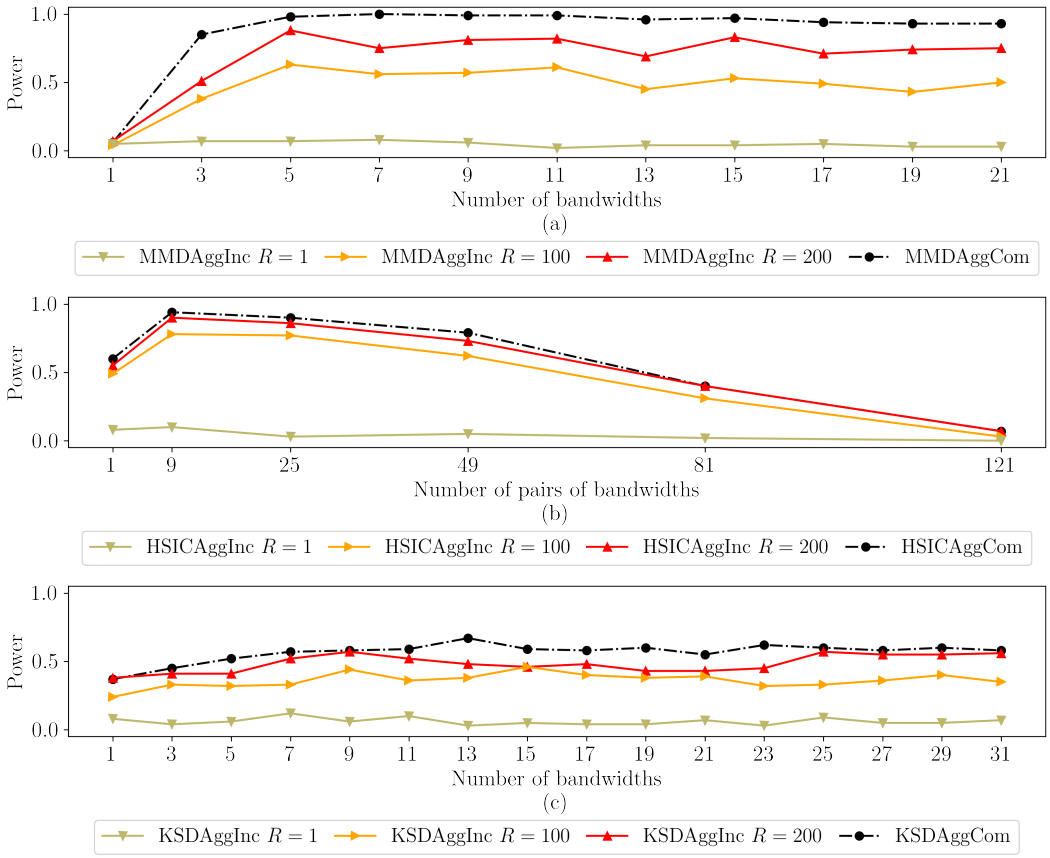

Figure 2: Increasing the collection of bandwidths in the setting of the experiments presented in Figure 1 for the two-sample *(a)*, independence *(b)*, and goodness-of-fit *(c)* problems. The power results are averaged over 100 repetitions.

D.3 MNIST Experiments

In Figure 3, we run experiments on the real-world MNIST dataset (LeCun et al., 2010) consisting of images of digits in dimension 784.

For the two sample problem, one distribution P consists of images of all digits and the other distribution is Q_i where Q_1 consists of images of only the five odd digits, Q_2 is Q_1 with 0, Q_3 is Q_2 with 2, Q_4 is Q_3 with 4, Q_5 is Q_4 with 6 (*i.e.* Q_5 consists of images of all digits expect 8). This setting has been previously considered by Schrab et al. (2021).

For the independence problem, we pair each image of a digit with the value of the digit. To make the problem more challenging, for some percentage of the data, we pair the images with values of random digits.

For the goodness-of-fit problem, the samples are drawn from the true MNIST dataset and the model is a Normalizing Flow (generative model which admits a density; Dinh et al., 2017; Kingma and Dhariwal, 2018) trained on the MNIST dataset. Since we have access only to pre-computed values of the score function evaluated at some MNIST samples and not to the score function itself, computing FSSD, L1 IMQ or Cauchy RFF is very challenging; for this reason the results for those tests are not reported.

Overall, we observe the same trends in Figure 3 for this high-dimensional real-world setting as we did in Figure 1 in the lower-dimensional setting in which the Sobolev smoothness assumption is satisfied for MMDAggInc and HSICAggInc. Indeed, the AggInc $R = 200$ tests clearly outperform the tests we compare against and even match the power of AggCom in three of the five settings considered.

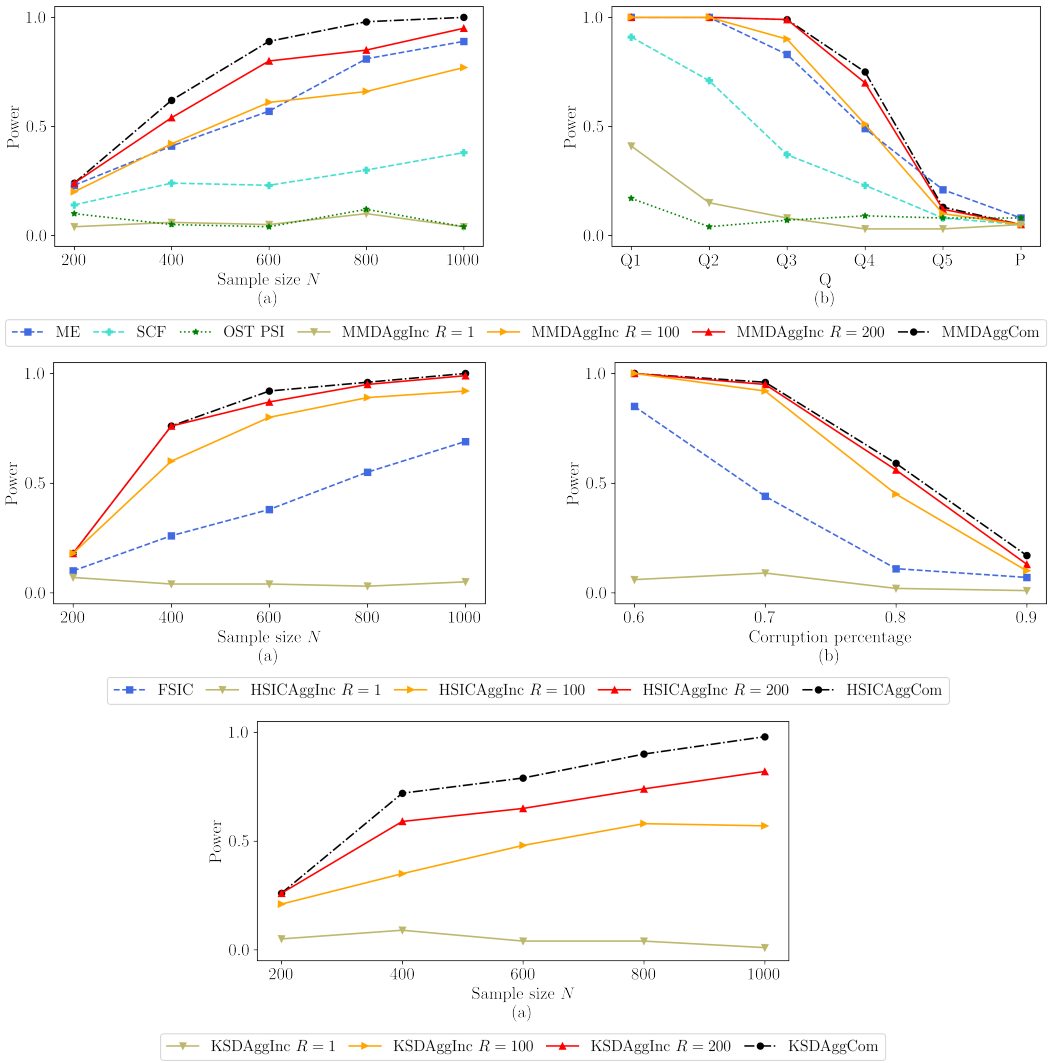

Figure 3: Experiments using the MNIST dataset for the two-sample *(a–b)*, independence *(c–d)*, and goodness-of-fit *(e)* problems. The power results are averaged over 100 repetitions.

ME obtains similar power to MMDAggInc $R = 100$ and even has higher power for $N = 800, 1000$ in Figure 3(a). We observe that HSICAggInc $R = 100$ significantly outperforms FSIC in both independence experiments in Figure 3(c,d). For the goodness-of-fit setting, the tests manage to detect that the true MNIST samples are not drawn from the density of the trained Normalizing Flow. There is a significant power difference between the four tests: KSDAggInc $R = 1, 100, 200$ and KSDAggCom.

D.4 Parameter choice for L1 IMQ and Cauchy RFF tests of Huggins and Mackey (2018)

As in the experiments section of Huggins and Mackey (2018) (and as for FSSD), 10 features are used when running Cauchy RFF and L1 IMQ. In their implementation for their experiments, they draw 5000 samples from the unnormalized density for covariance matrix estimation to simulate the null hypothesis (code: `RFDH0SimCovDrawV(n_draw=5000)`). This procedure causes long runtimes of roughly 16 seconds; this is much more computationally expensive than simulating the null using a wild bootstrap as KSDAggInc does.

We tried different values for `n_draw` and found that using `n_draw` $= 500$ has almost no effect on the test power and reduces the runtimes from 16 seconds for `n_draw` $= 5000$ to 2 seconds (as

reported in Figure 1(l)) for `n_draw` $= 500$. We tried smaller values than 500 for `n_draw` but this drastically decreased test power. We also verified that the test still has well-calibrated level when using `n_draw` $= 500$. We have used this tuned parameter `n_draw` $= 500$ in our experiments in Figure 1(i–l). In Figure 4, we show the power and runtime differences when using 500 or 5000 for `n_draw` for L1 IMQ and for Cauchy RFF for the setting considered in Figure 1(i–l), we also plot the test power and runtimes achieved by KSDAggInc $R = 200$.

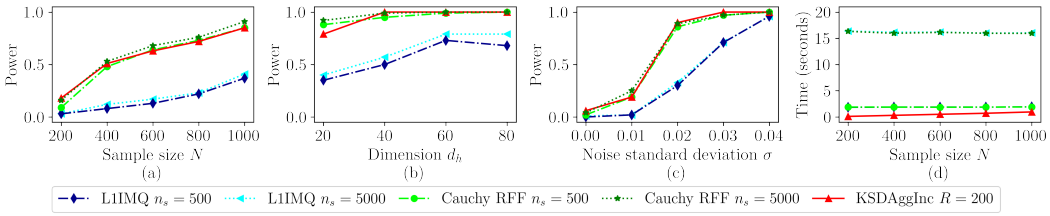

Figure 4: Goodness-of-fit experiment using a Gaussian-Bernoulli Restricted Boltzmann Machine. We consider L1 IMQ and Cauchy RFF tests of Huggins and Mackey (2018) drawing either $n_s =$ `n_draw` $= 5000$ or $n_s =$ `n_draw` $= 500$ samples to simulate the null hypothesis. The power results are averaged over 100 repetitions and the runtimes over 20 repetitions.

E Discussions

In this section, we provide detailed discussions on several subjects: the motivation behind Equations (8) and (9), incomplete MMD U-statistic with unbalanced sample sizes, comparison with related work, and future directions, in order.

E.1 Motivation behind expressions (8) and (9)

For the two-sample problem, Equation (26) of Kim et al. (2022, Section 6.1) gives an expression of the MMD U-statistic as

$$U_1 = \frac{1}{|\mathbf{i}_2^m||\mathbf{i}_2^n|} \sum_{(i,i') \in \mathbf{i}_2^m} \sum_{(j,j') \in \mathbf{i}_2^n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_j, Y_{j'}).$$

Now, one way to construct an incomplete MMD U-statistic would be to replace those two complete sums above with two incomplete sums (see Appendix E.2), but we do not want to take this approach in order to keep a unified framework across the three testing frameworks. We instead take the summation over $(i, i') \in \mathbf{i}_2^N$ and obtain an estimator

$$U_2 = \frac{1}{|\mathbf{i}_2^N|} \sum_{(i,i') \in \mathbf{i}_2^N} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_i, Y_{i'}),$$

where $N = \min(m, n)$. By assuming $N = m$, we denote by $\{L_1, \dots, L_n\}$ a random permutation of $\{1, \dots, n\}$. As noted by Kim et al. (2022, Section 6.1), the expectation of

$$\frac{1}{|\mathbf{i}_2^N|} \sum_{(i,i') \in \mathbf{i}_2^N} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_{L_i}, Y_{L_{i'}})$$

over $\{L_1, \dots, L_n\}$ is equal to U_1. This motivated our choice of incomplete MMD estimate in Equation (8) of our paper, which can be regarded as a generalization of U_2 above. Similarly, as discussed in Kim et al. (2022, Section 6.2), the complete HSIC U-statistic in Equation (3) can be viewed as the average of incomplete U-statistics. More specifically, let $\{L_1, \dots, L_{\lfloor N/2 \rfloor}\}$ be a $\lfloor N/2 \rfloor$-tuple uniformly sampled without replacement from $\{1, \dots, N\}$, and let $\{\tilde{L}_1, \dots, \tilde{L}_{\lfloor N/2 \rfloor}\}$ be another $\lfloor N/2 \rfloor$-tuple uniformly sampled without replacement from $\{1, \dots, N\} \setminus \{L_1, \dots, L_{\lfloor N/2 \rfloor}\}$. Then, the U-statistic in Equation (3) is the expectation of

$$\frac{1}{|\mathbf{i}_2^{\lfloor N/2 \rfloor}|} \sum_{(i_1,i_2) \in \mathbf{i}_2^{\lfloor N/2 \rfloor}} h_{k,\ell}^{\mathrm{HSIC}}(Z_{L_{i_1}}, Z_{L_{i_2}}; Z_{\tilde{L}_{i_1}}, Z_{\tilde{L}_{i_2}}),$$

over $\{L_1, \ldots, L_{\lfloor N/2 \rfloor}, \tilde{L}_1, \ldots, \tilde{L}_{\lfloor N/2 \rfloor}\}$. This motivates our incomplete HSIC estimate in Equation (9).

E.2 Incomplete MMD U-statistic with unbalanced sample sizes

Our incomplete U-statistic (8) for the two-sample problem is constructed using the minimum between m and n, and does not effectively incorporate the unbalanced sample sizes. When the sample sizes are highly unbalanced, one could also consider an alternative incomplete U-statistic given as

$$\begin{aligned} U_{\text{imb}} =& \frac{1}{|\mathcal{D}_m||\mathcal{D}_n|} \sum_{(i,j)\in\mathcal{D}_m} \sum_{(r,s)\in\mathcal{D}_n} h_k^{\text{MMD}}(X_i, X_j; Y_r, Y_s) \\ =& \frac{1}{|\mathcal{D}_m||\mathcal{D}_n|} \sum_{(i,j)\in\mathcal{D}_m} \sum_{(r,s)\in\mathcal{D}_n} \Big(k(X_i, X_j) - k(X_i, Y_s) - k(X_j, Y_r) + k(Y_r, Y_s) \Big). \end{aligned}$$

This expression, for example, results in a linear-time test for the choices $|\mathcal{D}_m| = c\sqrt{m}$ and $|\mathcal{D}_n| = c'\sqrt{n}$ for positive constants c and c' since $|\mathcal{D}_m||\mathcal{D}_n| = cc'\sqrt{m}\sqrt{n} \leq cc'\max(m, n)$. Other choices of design sizes are also possible to obtain linear-time tests. While this estimator is natural for the unbalanced scenario, the form of the test statistic does not allow us to use a wild bootstrap. Instead, one may need to rely on the permutation procedure to calibrate the test statistic, which leads to several theoretical and practical challenges explained below.

Theory. From a theoretical side, it is not too difficult to derive a variance bound (corresponding to Lemma 1) for the alternative estimate U_{imb}. However, deriving a quantile bound (corresponding to Lemma 2) for a permuted version of U_{imb} is highly non-trivial: the extension of the result of Kim et al. (2022, Theorem 6.3) to the case of the permuted version of U_{imb} is ongoing work.

Practice. Theoretically, the cost of computing B permuted estimates is $\mathcal{O}(B|\mathcal{D}_m||\mathcal{D}_n|)$ which would be the same as if we could use a wild bootstrap. However, in practice, the computational time will be much higher because for each permuted estimate, we need to evaluate the kernel matrix at new permuted pairs, while for the wild bootstrap, we do not need to compute any extra kernel values: this changes the computation times drastically. In order to avoid this, we would need to restrict ourselves to permutations for which we have already computed kernel values using the fact that $h_k^{\text{MMD}}(X_i, Y_s; Y_r, X_j) = -h_k^{\text{MMD}}(X_i, X_j; Y_r, Y_s)$. It remains as future work to study conditions under which the set of such permutations is larger than the set consisting of the identity only, and is also large enough to construct accurate quantiles.

E.3 Comparison with Li and Yuan (2019)

Li and Yuan (2019) also consider the three testing problems and study minimax optimality/adaptivity of their procedures over Sobolev balls. We now discuss and differentiate the approach by Li and Yuan (2019) from ours. First of all, their tests run in quadratic time and control the probability of type I error only asymptotically, while our proposed tests have well-calibrated non-asymptotic levels over a broader class of null distributions and are computationally efficient. Their theoretical guarantees hold only for the Gaussian kernel and with the smoothness restriction that $s > d/4$ while ours hold for a wide range of kernels (see Equation (13)) and for any $s > 0$ (see Theorem 2). We also point out that while they assume that both densities p and q lie in a Sobolev ball, we only require that their difference $p - q$ belongs to the Sobolev ball. Note that, they tackle the goodness-of-fit problem in a different way. They do not use the KSD and instead use a one-sample MMD with some expectations of the Gaussian kernel under the model. For a generic model density, one cannot compute such an expectation explicitly and hence cannot use their proposed test. In contrast, the KSD that we consider does not suffer from the same issue and it is more broadly applicable.

E.4 Potential future research directions

This subsection discusses potential directions for future work.

Beyond linear time tests. Potential directions for future work include studying the regime with $L \lesssim N$, which corresponds to "faster than linear" tests. For this sub-linear case, our results do not give a definite answer to the question as to whether the separation rate converges to zero. Future work

would focus on either deriving tighter bounds that converge to zero in this regime, or proving that the uniform separation rate is bounded below in this setting.

Computational and statistical trade-off. As shown in Theorem 1, the quadratic time MMD and HSIC tests are minimax rate optimal over Sobolev balls. To the best of our knowledge, it is unknown whether there exists a sub-quadratic time test that achieves the same rate optimality. Indeed, the current literature is mostly silent on optimising the power under computational constraints. Theorem 1(ii) demonstrates a trade-off between the computational budget and the separation rate focusing on incomplete U-statistics but our result does not tell us whether this trade-off is (universally) tight. We think this is one of the limitations of our work and hope that a follow-up study can make progress on this topic.

Continuously optimising a kernel without data splitting. In order to achieve a competitive power performance over a large class of alternatives, we combine finitely many kernels and construct an aggregated test. There is another line of work that considers a continuous collection of kernels (indexed by the bandwidth parameter on the positive line) and choose the kernel that maximises the (empirical) power. As far as we are aware, the current approach to continuously optimizing a kernel relies on data splitting, which often negatively affects the power performance. While our aggregated test does not require data splitting, in certain scenarios there might exist some important range of kernels not included in our finite collection which would lead to a more powerful test. It is therefore interesting to further extend our approach to the case of a continuous collection of kernels and investigate its statistical properties.

Improving conditions. Lastly, it would be interesting to see if the polynomial factor of β in our condition can be improved using a sharper concentration bound. Also, future work can be dedicated to figuring out whether the dependence of $\ln(\ln(N))$ in the adaptive rate can be improved. We leave these important and interesting questions to future work.

F Proofs

F.1 Proof of Proposition 1

The asymptotic level of the goodness-of-fit test using a wild bootstrap follows from the results of Shao (2010), Leucht and Neumann (2013), Chwialkowski et al. (2014, 2016). As pointed out by Schrab et al. (2022) for the complete U-statistics, the KSD test statistic and the wild bootstrapped KSD statistics are not exchangeable under the null, and hence non-asymptotic level cannot be proved using the result of Romano and Wolf (2005, Lemma 1).

The non-asymptotic level for the two-sample test follows exactly from the reasoning of Schrab et al. (2021, Propositon 1). The fact that we work with incomplete U-statistics rather than with their complete counterparts does not affect the proof of exchangeability of $U_\lambda^1, \ldots, U_\lambda^{B_1+1}$.

For the independence problem, Albert et al. (2022, Proposition 1) prove that the quadratic-time HSIC estimator and the permuted test statistics are exchangeable under the null hypothesis, it remains to be shown that this also holds in our incomplete setting using a wild bootstrap. Assuming that exchangeability under the null holds, the desired non-asymptotic level α can then be guaranteed using the result of Romano and Wolf (2005, Lemma 1), exactly as done by Albert et al. (2022, Proposition 1).

We now prove that $U_\lambda^1, \ldots, U_\lambda^{B_1+1}$ for the independence problem are exchangeable under the null. Since $U_\lambda^1, \ldots, U_\lambda^{B_1}$ are i.i.d. given the data, they are exchangeable under the null. So, we need to prove that

$$\sum_{(i,j)\in\mathcal{D}_{\lfloor N/2 \rfloor}} h_{k,\ell}^{\mathrm{HSIC}}\left(Z_i, Z_j, Z_{i+\lfloor N/2 \rfloor}, Z_{j+\lfloor N/2 \rfloor}\right) \tag{18}$$

is, under the null, distributed like

$$\sum_{(i,j)\in\mathcal{D}_{\lfloor N/2 \rfloor}} \epsilon_i \epsilon_j h_{k,\ell}^{\mathrm{HSIC}}\left(Z_i, Z_j, Z_{i+\lfloor N/2 \rfloor}, Z_{j+\lfloor N/2 \rfloor}\right) \tag{19}$$

where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. Rademacher random variables. Using the result of Schrab et al. (2021, Appendix B, Proposition 11), considering the identity $s_1(i) = i$ for $i = 1, \ldots, 2\lfloor N/2 \rfloor$, and the swap

function $s_{-1}(i) = i + \lfloor N/2 \rfloor$ and $s_{-1}(i + \lfloor N/2 \rfloor) = i$ for $i = 1, \dots, 2\lfloor N/2 \rfloor$, we have

$$
\begin{aligned}
&\sum_{(i,j)\in\mathcal{D}_{\lfloor N/2\rfloor}} \epsilon_i\epsilon_j h_{k,\ell}^{\mathrm{HSIC}}\big(Z_i, Z_j, Z_{i+\lfloor N/2\rfloor}, Z_{j+\lfloor N/2\rfloor}\big) \\
&= \frac{1}{4}\sum_{(i,j)\in\mathcal{D}_{\lfloor N/2\rfloor}} h_k^{\mathrm{MMD}}\big(X_i, X_j; X_{i+\lfloor N/2\rfloor}, X_{j+\lfloor N/2\rfloor}\big)\Big(\epsilon_i\epsilon_j h_\ell^{\mathrm{MMD}}\big(Y_i, Y_j; Y_{i+\lfloor N/2\rfloor}, Y_{j+\lfloor N/2\rfloor}\big)\Big) \\
&= \frac{1}{4}\sum_{\substack{(i,j)\in\\ \mathcal{D}_{\lfloor N/2\rfloor}}} h_k^{\mathrm{MMD}}\big(X_i, X_j; X_{i+\lfloor N/2\rfloor}, X_{j+\lfloor N/2\rfloor}\big) h_\ell^{\mathrm{MMD}}\big(Y_{s_{\epsilon_i}(i)}, Y_{s_{\epsilon_j}(j)}; Y_{s_{\epsilon_i}(i+\lfloor N/2\rfloor)}, Y_{s_{\epsilon_j}(j+\lfloor N/2\rfloor)}\big) \\
&= \big|\mathcal{D}_{\lfloor N/2\rfloor}\big|\,\overline{\mathrm{HSIC}}_{k,\ell}\big(\mathbb{Z}_N^\epsilon; \mathcal{D}_{\lfloor N/2\rfloor}\big)
\end{aligned}
$$

where $\mathbb{Z}_N^\epsilon := \Big(\big(X_i, Y_{s_{\epsilon_i}(i)}\big)\Big)_{1\le i\le N}$ with $Y_{s_{\epsilon_i}(i)} \in \{Y_i, Y_{i+\lfloor N/2\rfloor}\}$ for $i = 1, \dots, \lfloor N/2 \rfloor$ and $Y_{s_{\epsilon_i}(i)} \in \{Y_i, Y_{i-\lfloor N/2\rfloor}\}$ for $i = 1 + \lfloor N/2 \rfloor, \dots, 2\lfloor N/2 \rfloor$. Now, under the null the variables $(X_i)_{1\le i\le N}$ and $(Y_i)_{1\le i\le N}$ are independent, so $\mathbb{Z}_N^\epsilon$ is distributed like $\mathbb{Z}_N$. We deduce that $\overline{\mathrm{HSIC}}_{k,\ell}\big(\mathbb{Z}_N; \mathcal{D}_{\lfloor N/2\rfloor}\big)$ and $\overline{\mathrm{HSIC}}_{k,\ell}\big(\mathbb{Z}_N^\epsilon; \mathcal{D}_{\lfloor N/2\rfloor}\big)$ have the same distribution under the null, and hence, that the terms in Equations (18) and (19) also have the same distribution under the null. We deduce that $U_\lambda^1, \dots, U_\lambda^{B_1+1}$ for the independence problem are exchangeable under the null, which completes the proof.

F.2 Proof of Lemma 1

Consider the case of fixed design. Using the variance expression of Lee (1990, Theorem 2, p. 190), we have

$$\mathrm{var}\big(\overline{U}\big) = \frac{f_1\sigma_1^2 + f_2\sigma_2^2}{|\mathcal{D}|^2}$$

where f_i is the number of pairs of sets in the design $\mathcal{D}$ that have i elements in common. The pairs of sets in $\mathcal{D}$ with 2 elements in common are $\{(\{i,j\},\{i,j\}) : (i,j) \in \mathcal{D}\}$, so $f_2 = |\mathcal{D}|$. We now calculate the number of pairs of sets in $\mathcal{D}$ with 1 element in common. We start with a pair $(i,j) \in \mathcal{D}$ (there are $|\mathcal{D}|$ such pairs). The number of pairs in $\mathcal{D}$ which have one element in common with (i,j) is upper bounded by the number of pairs in $\mathbf{i}_2^N$ which have one element in common with (i,j), those are $\{\{i,r\} : 1 \le r \le N, r \ne i\} \cup \{\{j,r\} : 1 \le r \le N, r \ne j\}$ of size smaller than $2N$. We deduce that $f_1 \le 2N|\mathcal{D}|$. Combining those results, we obtain

$$\mathrm{var}\big(\overline{U}\big) \le \frac{f_1\sigma_1^2 + f_2\sigma_2^2}{|\mathcal{D}|^2} \le \frac{2N}{|\mathcal{D}|}\sigma_1^2 + \frac{1}{|\mathcal{D}|}\sigma_2^2$$

as desired.

Let us now consider the random design case. Recall that using the variance expression of the complete U-statistic U of Lee (1990, Theorem 3 p. 12), we can obtain that

$$\mathrm{var}(U) \lesssim \frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N^2}$$

as done by Kim et al. (2022, Appendix E) and Albert et al. (2022, Lemma 10). Using the result of Lee (1990, Theorem 4 p. 193), the variance of the incomplete U-statistic $\overline{U}$ can be expressed in terms of the variance of the complete U-statistic U. For random design with replacement, we have

$$
\begin{aligned}
\mathrm{var}\big(\overline{U}\big) &= \frac{\sigma_2^2}{|\mathcal{D}|} + \left(1 - \frac{1}{|\mathcal{D}|}\right)\mathrm{var}(U) \\
&\lesssim \frac{\sigma_1^2}{N} + \left(\frac{1}{|\mathcal{D}|} + \frac{1}{N^2}\right)\sigma_2^2 \\
&\lesssim \frac{\sigma_1^2}{N} + \frac{1}{|\mathcal{D}|}\sigma_2^2.
\end{aligned}
$$

Letting $S := N(N-1)/2$, for random design without replacement, we have

$$\begin{aligned}\mathrm{var}\big(\overline{U}\big) &= \frac{S-|\mathcal{D}|}{|\mathcal{D}|(S-1)}\sigma_2^2 + \frac{S}{S-1}\left(1-\frac{1}{|\mathcal{D}|}\right)\mathrm{var}(U) \\ &\lesssim \frac{\sigma_1^2}{N} + \left(\frac{1}{|\mathcal{D}|}+\frac{1}{N^2}\right)\sigma_2^2 \\ &\lesssim \frac{\sigma_1^2}{N} + \frac{1}{|\mathcal{D}|}\sigma_2^2.\end{aligned}$$

We have used the fact that $\frac{1}{N^2} \leq \frac{1}{|\mathcal{D}|}$ since $|\mathcal{D}| \leq N^2$.

F.3 Proof of Lemma 2

We rely on the concentration bound for i.i.d. Rademacher chaos of de la Peña and Giné (1999, Corollary 3.2.6) which as presented in Kim et al. (2022, Equation 39) takes the form

$$\mathbb{P}_\epsilon\left(\left|\sum_{(i,j)\in \mathbf{i}_2^N} \epsilon_i\epsilon_j a_{i,j}\right| \geq t\right) \leq 2\exp\left(-Ct\left(\sum_{(i,j)\in \mathbf{i}_2^N} a_{i,j}^2\right)^{-1}\right)$$

for some constant $C > 0$ and for every $t \geq 0$, where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. Rademacher random variables taking values in $\{-1, 1\}$. Letting

$$a_{i,j} := \frac{h(Z_i, Z_j)}{|\mathcal{D}|}\mathbf{1}\Big[(i,j)\in\mathcal{D}\Big] \qquad \text{for} \qquad (i,j)\in \mathbf{i}_2^N,$$

where $\mathbf{1}$ denotes the indicator function, we obtain

$$\begin{aligned}\mathbb{P}_\epsilon\left(\frac{1}{|\mathcal{D}|}\left|\sum_{(i,j)\in\mathcal{D}} \epsilon_i\epsilon_j h(Z_i,Z_j)\right| \geq t \;\middle|\; \mathbb{Z}_N, \mathcal{D}\right) &\leq 2\exp\left(-Ct\left(\frac{1}{|\mathcal{D}|^2}\sum_{(i,j)\in\mathcal{D}} h(Z_i,Z_j)^2\right)^{-1}\right) \\ &\leq 2\exp\left(-Ct\left(\frac{1}{|\mathcal{D}|^2}\sum_{(i,j)\in\mathbf{i}_2^N} h(Z_i,Z_j)^2\right)^{-1}\right)\end{aligned}$$

which concludes the proof.

F.4 Proof of Theorem 1

We start by reviewing the steps of proofs of Albert et al. (2022) and Schrab et al. (2021, 2022) who prove that, for each of the three respective testing frameworks, a sufficient condition to ensure control of the probability of type II error for the quadratic-time test is the existence of a constant $C > 0$ such that

$$\|p-q\|_2^2 \;\geq\; \|(p-q) - T_\lambda(p-q)\|_2^2 + C\frac{1}{N}\frac{\ln(1/\alpha)}{\beta}\sigma_{2,\lambda}. \tag{20}$$

Those quadratic-time tests use the complete U-statistics defined in Equations (1), (3) and (5), which we denote as U_λ. The key results for their proofs rely on deriving variance and quantile bounds.

The variance bound is of the form

$$\mathrm{var}(U_\lambda) \lesssim \frac{1}{N}\sigma_1^2 + \frac{1}{N^2}\sigma_2^2 \tag{21}$$

where they show that for $h_\lambda \in \big\{h_{k_\lambda}^{\mathrm{MMD}}, h_{k_\lambda,\ell_\mu}^{\mathrm{HSIC}}, h_{k_\lambda,p}^{\mathrm{KSD}}\big\}$ defined in Equations (2), (4) and (6)

$$\sigma_{1,\lambda}^2 := \mathrm{var}\big(\mathbb{E}\big[h_\lambda(Z,Z')\big|Z'\big]\big) \lesssim \big\|T_\lambda(p-q)\big\|_2^2$$

and

$$\sigma_{2,\lambda}^2 := \mathrm{var}(h_\lambda(Z,Z')) = \mathbb{E}\big[h_\lambda(Z,Z')^2\big] \lesssim \frac{1}{\lambda_1\cdots\lambda_d} \tag{22}$$

for some $C > 0$, where the last inequality holds only for $h_\lambda \in \{h^{\mathrm{MMD}}_{k_\lambda}, h^{\mathrm{HSIC}}_{k_\lambda,\ell_\mu}\}$.

The quantile bound (Schrab et al., 2021, Proposition 4) is of the form

$$\mathbb{P}\left(\widehat{q}^{\lambda,U,B_1}_{1-\alpha} \lesssim \frac{1}{N}\frac{1}{\sqrt{\delta}}\ln\left(\frac{1}{\alpha}\right)\sigma_{2,\lambda}\right) \geq 1-\delta$$

for $\delta \in (0,1)$, where $\widehat{q}^{\lambda,U,B_1}_{1-\alpha}$ is the quantile obtained using B_1 wild bootstrapped similarly to the one defined in Equation (14) but using the complete U-statistic. Relying on Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990), Schrab et al. (2021, Proposition 4) show that it suffices to prove the bound

$$\mathbb{P}\left(\widehat{q}^{\lambda,U,\infty}_{1-\alpha} \lesssim \frac{1}{N}\frac{1}{\sqrt{\delta}}\ln\left(\frac{1}{\alpha}\right)\sigma_{2,\lambda}\right) \geq 1-\delta \tag{23}$$

for the true wild bootstrap quantile $\widehat{q}^{\lambda,U,\infty}_{1-\alpha}$ without finite approximation.

Combining those variance and quantile bounds using Chebyshev's inequality (Chebyshev, 1899), they obtain a condition guaranteeing power in terms of the MMD, HSIC and KSD. By expressing these three measures as an RHKS inner product

$$\langle p-q, T_\lambda(p-q)\rangle = \frac{1}{2}\Big(\|p-q\|_2^2 + \|T_\lambda(p-q)\|_2^2 - \|(p-q)-T_\lambda(p-q)\|_2^2\Big),$$

they obtain the condition in Equation (20) which guarantees high power in terms of $\|p-q\|_2^2$. Albert et al. (2022) and Schrab et al. (2021) then derive the minimax rate $(1/N)^{2s/(4s+d)}$ over the Sobolev ball $\mathcal{S}^s_d(R)$ for the independence and two-sample tests using the bandwidths $\lambda^*_i := (1/N)^{2/(4s+d)}$ for $i = 1,\ldots,d$.

For Theorem 1, we need to obtain the condition in Equation (20) with $1/N$ replaced by N/L. Hence, following their reasoning, in order to prove Theorem 1 (i) & (ii), it suffices to derive variance and quantiles bounds for incomplete U-statistics which have the form of Equations (21) and (23) with $1/N$ replaced by N/L, which we now do.

Using the variance bound for incomplete U-statistics $\overline{U}_\lambda$ of Lemma 1, together with the fact that the design size $L := |\mathcal{D}|$ is smaller than N^2 so that $1/L = L/L^2 \leq N^2/L^2$, we obtain for fixed design that

$$\mathrm{var}(\overline{U}_\lambda) \lesssim \frac{N}{L}\sigma^2_{1,\lambda} + \frac{1}{L}\sigma^2_{2,\lambda} \lesssim \frac{N}{L}\sigma^2_{1,\lambda} + \left(\frac{N}{L}\right)^2\sigma^2_{2,\lambda}.$$

We get the same bound for random design since

$$\mathrm{var}(\overline{U}_\lambda) \lesssim \frac{1}{N}\sigma^2_{1,\lambda} + \frac{1}{L}\sigma^2_{2,\lambda} \lesssim \frac{N}{L}\sigma^2_{1,\lambda} + \left(\frac{N}{L}\right)^2\sigma^2_{2,\lambda},$$

as desired.

For the quantile bound, we use Lemma 2 which, for $A^2_\lambda := L^{-2}\sum_{(i,j)\in \mathbf{i}^N_2} h_\lambda(Z_i,Z_j)^2$, gives that there exists some[9] $C > 0$ such that

$$\mathbb{P}_\epsilon\left(\overline{U}^\epsilon_\lambda \geq t \,\middle|\, \mathbb{Z}_N, \mathcal{D}\right) \leq \mathbb{P}_\epsilon\left(\left|\overline{U}^\epsilon_\lambda\right| \geq t \,\middle|\, \mathbb{Z}_N, \mathcal{D}\right) \leq 2\exp\left(-\frac{t}{CA_\lambda}\right).$$

Setting $\alpha := 2\exp(-t/CA_\lambda)$, we obtain

$$\widehat{q}^{\lambda,\overline{U},\infty}_{1-\alpha} = t = CA_\lambda \ln(2/\alpha).$$

For $\delta \in (0,1)$, using Markov's inequality, we obtain

$$\mathbb{P}\left(A^2_\lambda \leq \frac{1}{\delta}\mathbb{E}\left[A^2_\lambda\right]\right) \geq 1-\delta$$

[9] For simplicity, we work with $C^{-1} > 0$ rather than with $C > 0$.

where

$$\mathbb{E}\big[A_\lambda^2\big] = \mathbb{E}\left[\frac{1}{L^2}\sum_{(i,j)\in\mathbf{i}_2^N} h(Z_i,Z_j)^2\right] = \frac{N(N-1)}{L^2}\mathbb{E}\big[h_\lambda(Z,Z')^2\big] \lesssim \frac{N^2}{L^2}\sigma_{2,\lambda}^2$$

using Equation (22). We deduce that

$$\begin{aligned}
1-\delta &\leq \mathbb{P}\left(A_\lambda^2 \leq \frac{1}{\delta}\,\mathbb{E}\big[A_\lambda^2\big]\right)\\
&= \mathbb{P}\left(\widehat{q}_{1-\alpha}^{\lambda,\overline{U},\infty} \leq C\frac{1}{\sqrt{\delta}}\ln\left(\frac{2}{\alpha}\right)\sqrt{\mathbb{E}[A_\lambda^2]}\right)\\
&\leq \mathbb{P}\left(\widehat{q}_{1-\alpha}^{\lambda,\overline{U},\infty} \lesssim \frac{1}{\sqrt{\delta}}\frac{N}{L}\ln\left(\frac{1}{\alpha}\right)\sigma_{2,\lambda}\right)
\end{aligned}$$

where we absorbed the constant C in the notation '$\lesssim$', and where we used the fact that $\ln(2/\alpha) \lesssim \ln(1/\alpha)$ since $\alpha \in (0, e^{-1})$. This concludes the proof.

F.5 Proof of Theorem 3

F.5.1 Proof of Theorem 3 (i)

In this setting, we consider as proposed by Kim et al. (2022, Equation 32) the permuted HSIC complete U-statistic

$$U_N^\pi := \frac{1}{\left|\mathbf{i}_4^N\right|}\sum_{(i,j,r,s)\in\mathbf{i}_4^N} h_{k,\ell}^{\mathrm{HSIC}}\left((X_i,Y_{\pi_i}),(X_j,Y_{\pi_j}),(X_r,Y_{\pi_r}),(X_s,Y_{\pi_s})\right)$$

for a permutation π of the indices $\{1,\ldots,N\}$, and for $h_{k,\ell}^{\mathrm{HSIC}}$ as defined in Equation (4).

Applying the exponential concentration bound of Kim et al. (2022, Theorem 6.3), which uses the result of de la Peña and Giné (1999, Theorem 4.1.12), we obtain that there exist constants $C_1, C_2 > 0$ such that

$$\mathbb{P}_\pi(U_N^\pi \geq t \mid \mathbb{Z}_N) \leq C_1 \exp\left(-C_2 \min\left(\frac{Nt}{\Lambda_N}, \frac{Nt^{2/3}}{M_N^{2/3}}\right)\right) \tag{24}$$

where

$$\Lambda_N^2 := \frac{1}{N^4}\sum_{i=1}^N\sum_{j=1}^N\sum_{r=1}^N\sum_{s=1}^N k_\lambda(X_i,X_j)^2\ell_\mu(Y_r,Y_s)^2$$

and

$$\begin{aligned}
M_N &:= \max_{1\leq i,j,r,s\leq N}\left|k_\lambda(X_i,X_j)\ell_\mu(Y_r,Y_s)\right|\\
&= \max_{1\leq i,j,r,s\leq N}\left|\prod_{a=1}^{d_x}\frac{1}{\lambda_a}K_a\left(\frac{(X_i)_a-(X_j)_a}{\lambda_a}\right)\prod_{b=1}^{d_y}\frac{1}{\lambda_b}L_b\left(\frac{(Y_r)_b-(Y_s)_b}{\lambda_b}\right)\right|\\
&\lesssim \frac{1}{\lambda_1\ldots\lambda_{d_x}\mu_1\ldots\mu_{d_y}}\\
&= \frac{1}{\lambda_1\cdots\lambda_d}
\end{aligned}$$

since the functions $K_1,\ldots,K_{d_x}$ and $L_1,\ldots,L_{d_y}$ are bounded, and where we recall our notational convention that $d := d_x + d_y$ and $\lambda_{d_x+i} := \mu_i$ for $i = 1,\ldots,d_y$.

Using the reasoning of Schrab et al. (2021, Proposition 3), we see that the results of Albert et al. (2022, Equations C.17, C.18 & C.19) hold not only for the Gaussian kernel but more generally for

any kernels of the form of Equation (13). Those results give us that

$$\mathbb{E}\big[k_\lambda(X_1,X_2)^2\ell_\mu(Y_1,Y_2)^2\big] \lesssim \frac{1}{\lambda_1\dots\lambda_{d_x}\mu_1\dots\mu_{d_y}} = \frac{C}{\lambda_1\cdots\lambda_d},$$
$$\mathbb{E}\big[k_\lambda(X_1,X_2)^2\ell_\mu(Y_1,Y_3)^2\big] \lesssim \frac{1}{\lambda_1\cdots\lambda_d},$$
$$\mathbb{E}\big[k_\lambda(X_1,X_2)^2\ell_\mu(Y_3,Y_4)^2\big] \lesssim \frac{1}{\lambda_1\cdots\lambda_d},$$

the constant in the notation '$\lesssim$' depends only on d and M, where we recall that $\max\left(\|p\|_\infty, \|p\|_\infty\right) \leq M$ by assumption. We deduce that

$$\mathbb{E}\big[\Lambda_N^2\big] \lesssim \frac{1}{\lambda_1\cdots\lambda_d}.$$

As explained in Appendix F.4, by relying on Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990) as done by Schrab et al. (2021, Proposition 4), it is sufficient to prove upper bounds for the true permutation quantile $\widehat{q}_{1-\alpha}^{\lambda,\infty}$ without finite approximation. From Equation (24), we obtain that this quantile satisfies

$$\widehat{q}_{1-\alpha}^{\lambda,\infty} \lesssim \max\left(\frac{\Lambda_N}{N}\ln\left(\frac{1}{\alpha}\right), \frac{M_N}{N^{3/2}}\ln\left(\frac{1}{\alpha}\right)^{3/2}\right).$$

Using Markov's inequality and bounds obtained above, we get that

$$\widehat{q}_{1-\alpha}^{\lambda,\infty} \lesssim \max\left(\frac{\sqrt{\mathbb{E}\left[\Lambda_N^2\right]}}{\sqrt{\delta}N}\ln\left(\frac{1}{\alpha}\right), \frac{M_N}{N^{3/2}}\ln\left(\frac{1}{\alpha}\right)^{3/2}\right)$$
$$\widehat{q}_{1-\alpha}^{\lambda,\infty} \lesssim \max\left(\frac{\ln(1/\alpha)}{\sqrt{\delta}N\sqrt{\lambda_1\cdots\lambda_d}}, \frac{\ln(1/\alpha)^{3/2}}{N^{3/2}\lambda_1\cdots\lambda_d}\right) \tag{25}$$

holds with probability at least $1-\delta$ where $\delta \in (0,1)$.

Now, recall that by assumption $4s > d$ and $\lambda_i^* = N^{-2/(4s+d)}$ for $i = 1,\dots,d$, so that

$$\frac{1}{\lambda_1^*\cdots\lambda_d^*} = N^{2d/(4s+d)} < N \qquad \Longleftrightarrow \qquad N^{-1/2} < \sqrt{\lambda_1^*\cdots\lambda_d^*}$$

which gives

$$\widehat{q}_{1-\alpha}^{\lambda^*,\infty} \lesssim \max\left(\frac{\ln(1/\alpha)}{\sqrt{\delta}N\sqrt{\lambda_1^*\cdots\lambda_d^*}}, \frac{\ln(1/\alpha)^{3/2}}{N\sqrt{\lambda_1^*\cdots\lambda_d^*}}\right)$$
$$\lesssim \frac{\ln(1/\alpha)^{3/2}}{\sqrt{\delta}N\sqrt{\lambda_1^*\cdots\lambda_d^*}}$$

holding with probability at least $1-\delta$, since $\alpha \in \left(0, e^{-1}\right)$. By combining this result with the reasoning of Albert et al. (2022) as explained in Appendix F.4, we obtain that the probability of type II error of the test is controlled by $\beta \in (0,1)$ when

$$\|p-q\|_2^2 \;\geq\; \|(p-q) - T_{\lambda^*}(p-q)\|_2^2 + C\frac{1}{N}\frac{\ln(1/\alpha)^{3/2}}{\beta\sqrt{\lambda_1^*\cdots\lambda_d^*}}.$$

We have recovered the correct dependency with respect to N and λ with an improved α-dependency of $\ln(1/\alpha)^{3/2}$ compared to the $\alpha^{-1/2}$ dependency obtained by Kim et al. (2022, Proposition 8.7). The proof of minimax optimality of the quadratic-time test with fixed bandwidth λ^* does not depend on the α-dependency and can be derived in both our setting and the one of Kim et al. (2022) using quantiles obtained from permutations by following the reasoning of Albert et al. (2022, Corollary 2). We obtain that the uniform separation rate over the Sobolev ball $\mathcal{S}_d^s(R)$ is, up to a constant, $(1/N)^{2s/(4s+d)}$. The improved α-dependency is crucial for deriving the rate of the aggregated quadratic-time test over Sobolev balls because the weights appear in the α-term (i.e. α is replaced by αw_λ which depends on the sample size N).

F.5.2 Proof of Theorem 3 (ii)

Similarly to the proofs of Albert et al. (2022, Corollary 3) and Schrab et al. (2021, Corollary 10), consider

$$\ell^* := \left\lceil \frac{2}{4s+d} \log_2\left(\frac{N}{\ln(\ln(N))}\right)\right\rceil \leq \left\lceil \frac{2}{d} \log_2\left(\frac{N}{\ln(\ln(N))}\right)\right\rceil$$

and the bandwidth $\lambda^* := \left(2^{-\ell^*}, \ldots, 2^{-\ell^*}\right) \in \Lambda$ which satisfies

$$\ln\left(\frac{1}{w_{\lambda^*}}\right) \lesssim \ln\left(\ell^*\right) \lesssim \ln(\ln(N))$$

as $w_{\lambda^*} := 6\pi^{-2}\left(\ell^*\right)^{-2}$, and

$$\frac{1}{2}\left(\frac{\ln(\ln(N))}{N}\right)^{2/(4s+d)} \leq \lambda_i^* \leq \left(\frac{\ln(\ln(N))}{N}\right)^{2/(4s+d)}$$

for $i = 1, \ldots, d$. Since $4s > d$, we have

$$\frac{1}{\lambda_1^* \cdots \lambda_d^*} \leq 2^d \left(\frac{N}{\ln(\ln(N))}\right)^{\frac{2d}{4s+d}} \lesssim \frac{N}{\ln(\ln(N))} \quad \Longleftrightarrow \quad N^{-1/2} \lesssim \frac{\sqrt{\lambda_1^* \cdots \lambda_d^*}}{\sqrt{\ln(\ln(N))}},$$

all with different constants. By Equation (25), we get that

$$\widehat{q}^{\lambda^*,\infty}_{1-\alpha w_{\lambda^*}} \lesssim \max\left(\frac{\ln\big(1/(\alpha w_{\lambda^*})\big)}{\sqrt{\delta} N \sqrt{\lambda_1^* \cdots \lambda_d^*}}, \frac{\ln\big(1/(\alpha w_{\lambda^*})\big)^{3/2}}{N^{3/2} \lambda_1^* \cdots \lambda_d^*}\right) \tag{26}$$

holds with probability at least $1 - \delta$ for $\delta \in (0, 1)$. If the largest term is the first one, then we get

$$\begin{aligned} \widehat{q}^{\lambda^*,\infty}_{1-\alpha w_{\lambda^*}} &\lesssim \frac{\ln(1/\alpha) + \ln(1/w_{\lambda^*}))}{\sqrt{\delta} N \sqrt{\lambda_1^* \cdots \lambda_d^*}}, \\ \widehat{q}^{\lambda^*,\infty}_{1-\alpha w_{\lambda^*}} &\lesssim \frac{\ln(\ln(N))}{\sqrt{\delta} N \sqrt{\lambda_1^* \cdots \lambda_d^*}}, \end{aligned} \tag{27}$$

where the constant also depends on α. The result then follows exactly as in the proofs of Albert et al. (2022, Corollary 3) and Schrab et al. (2021, Corollary 10). So, we consider the case where the second term is the largest one, so that

$$\begin{aligned} \widehat{q}^{\lambda^*,\infty}_{1-\alpha w_{\lambda^*}} &\lesssim \frac{\ln\big(1/(\alpha w_{\lambda^*})\big)^{3/2}}{N^{3/2} \lambda_1^* \cdots \lambda_d^*} \\ &\lesssim \frac{\ln\big(1/w_{\lambda^*}\big)^{3/2}}{N \lambda_1^* \cdots \lambda_d^*} N^{-1/2} \\ &\lesssim \frac{\ln(\ln(N))^{3/2}}{N \lambda_1^* \cdots \lambda_d^*} \frac{\sqrt{\lambda_1^* \cdots \lambda_d^*}}{\sqrt{\ln(\ln(N))}} \\ &= \frac{\ln(\ln(N))}{N \sqrt{\lambda_1^* \cdots \lambda_d^*}}. \end{aligned}$$

We have recovered the same dependency as in Equation (27) when considering the first term as the largest one, and the proof then follows exactly the ones of Albert et al. (2022, Corollary 3) and Schrab et al. (2021, Corollary 10). We have treated both cases in Equation (26), we conclude that the uniform separation rate over the Sobolev balls $\left\{\mathcal{S}_d^s(R) : R > 0, s > d/4\right\}$ of the quadratic-time aggregated test using a quantile obtained with permutations is (up to a constant)

$$\left(\frac{\ln(\ln(N))}{N}\right)^{2s/(4s+d)}.$$