Computationally Efficient Aggregated Kernel Tests using Incomplete U-statistics

Anonymous Author(s)
Affiliation
Address
email

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. For the test thresholds, we derive a quantile bound for wild bootstrapped incomplete U-statistics, which is of independent interest. We derive 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 is known to lead to state-of-the-art powerful tests (Fromont et al., 2012, 2013; Albert et al., 2022; Schrab et al., 2021, 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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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: linear-time 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), the current state-of-the-art 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) and Huggins and Mackey (2018) on quadratic-time tests, and of Ho and Shieh (2006), Zaremba et al. (2013) and Zhang et al. (2018) on the use of block U-statistics which have 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 in linear time, 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 of linear-time tests using incomplete U-statistics for a fixed kernel bandwidth, in Section 5. In particular, we obtain 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 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://anonymous.4open.science/r/agginc-10EF/README.md.

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.

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 so that there exists a constant[2] $C > 0$ such that $\max(m, n) \leq C \min(m, n)$. 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

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

[2] We use the convention that all constants are generically denoted by C, even though they are different.

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 $\{1, \ldots, b\}$ 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

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

Independence testing. In this problem, we have access to i.i.d. pairs of samples $\mathbb{Z}_N \coloneqq \big(Z_i\big)_{1\le i\le N} = \big((X_i, Y_i)\big)_{1\le i\le N}$ 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 $z_a = (x_a, y_a) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $a = 1, \ldots, 4$, we let

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

Goodness-of-fit testing. For this problem, we are given a model density p on $\mathbb{R}^d$ and i.i.d. samples $\mathbb{Z}_N \coloneqq (Z_i)_{1\le i\le N}$ 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,

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

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

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

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 $\mathbb{E}_q\Big[\big\|\nabla \log \frac{p(Z)}{q(Z)}\big\|_2^2\Big] < \infty$.

Quantile estimation. Multiple strategies have been proposed to estimate the quantiles of test statistics under the null for the 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 A.

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 linear-time 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),

$$\overline{\mathrm{MMD}}_k^2\big(\mathbb{X}_m, \mathbb{Y}_n; \mathcal{D}_N\big) \coloneqq \frac{1}{\big|\mathcal{D}_N\big|} \sum_{(i,j)\in\mathcal{D}_N} h_k^{\mathrm{MMD}}(X_i, X_j; Y_i, Y_j), \tag{7}$$

$$\overline{\mathrm{HSIC}}_{k,\ell}\big(\mathbb{Z}_N; \mathcal{D}_{\lfloor N/2\rfloor}\big) \coloneqq \frac{1}{\big|\mathcal{D}_{\lfloor N/2\rfloor}\big|} \sum_{(i,j)\in\mathcal{D}_{\lfloor N/2\rfloor}} h_{k,\ell}^{\mathrm{HSIC}}\left(Z_i, Z_j, Z_{i+\lfloor N/2\rfloor}, Z_{j+\lfloor N/2\rfloor}\right), \tag{8}$$

$$\overline{\mathrm{KSD}}_{p,k}^2\big(\mathbb{Z}_N; \mathcal{D}_N\big) \coloneqq \frac{1}{\big|\mathcal{D}_N\big|} \sum_{(i,j)\in\mathcal{D}_N} h_{k,p}^{\mathrm{KSD}}(Z_i, Z_j), \tag{9}$$

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 (7) to (9) become random quantities given the data. The results presented in this paper hold for both deterministic and random (without replacement) design choices. By fixing the design sizes in Equations (7) to (9) to be

$$\big|\mathcal{D}_N\big| = \big|\mathcal{D}_{\lfloor N/2\rfloor}\big| = cN \tag{10}$$

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 (7) to (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.

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 $\sigma_1^2 \coloneqq \mathrm{var}\big(\mathbb{E}\big[h(Z,Z')\big|Z'\big]\big)$ and $\sigma_2^2 \coloneqq \mathrm{var}(h(Z,Z'))$ with different bounds depending on the design choice. For deterministic design $\mathcal{D}_d$, and for random design $\mathcal{D}_r$, we have*

$$\mathrm{var}\big(\overline{U}\big) \le C\left(\frac{N}{|\mathcal{D}_d|}\sigma_1^2 + \frac{1}{|\mathcal{D}_d|}\sigma_2^2\right) \qquad \textit{and} \qquad \mathrm{var}\big(\overline{U}\big) \le C\left(\frac{1}{N}\sigma_1^2 + \left(\frac{1}{|\mathcal{D}_r|} + \frac{1}{N^2}\right)\sigma_2^2\right).$$

The proof of Lemma 1 is deferred to Appendix D. 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 $N \leq |\mathcal{D}| \leq N^2$, the variance bound is $\sigma_1^2/N + \sigma_2^2/|\mathcal{D}|$, and if $|\mathcal{D}| \leq N$ 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

$$\overline{U}^{\epsilon}\left(\mathbb{Z}_N; \mathcal{D}\right) \coloneqq \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} \epsilon_i \epsilon_j h(Z_i, Z_j) \tag{11}$$

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*

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

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

Lemma 2 is proved in Appendix E. 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]

$$k_\lambda(x, y) \coloneqq \prod_{i=1}^{d_x} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right), \qquad \ell_\mu(x, y) \coloneqq \prod_{i=1}^{d_y} \frac{1}{\mu_i} L_i\left(\frac{x_i - y_i}{\mu_i}\right), \tag{12}$$

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 problem, we work with the two kernels k_λ and ℓ_μ, and for ease of notation we let $d \coloneqq d_x + d_y$ and $\lambda_{d_x+i} \coloneqq \mu_i$ for $i = 1, \ldots, d_y$. We also simply write $p \coloneqq p_{xy}$ and $q \coloneqq p_x \otimes p_y$. 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 (7) to (9) by

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

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 $U_\lambda^{B_1+1} \coloneqq \overline{U}_\lambda$ together

[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).

with B_1 wild bootstrapped incomplete U-statistics $U_\lambda^1, \dots, U_\lambda^{B_1}$ computed as in Equation (11), and estimating the $(1-\alpha)$-quantile with a Monte Carlo approximation

$$\widehat{q}_{1-\alpha}^{\lambda} := \inf\left\{ t \in \mathbb{R} : 1-\alpha \leq \frac{1}{B_1+1} \sum_{b=1}^{B_1+1} \mathbf{1}\big(U_\lambda^b \leq t\big) \right\} = U_\lambda^{\bullet \lceil B_1(1-\alpha) \rceil}, \tag{13}$$

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

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

We show in Proposition 1 that the test Δ_α^λ 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 C).

Proposition 1. *The test Δ_α^λ has level $\alpha \in (0,1)$, i.e., $\mathbb{P}_{\mathcal{H}_0}\big(\Delta_\alpha^\lambda(\mathbb{Z}_N) = 1\big) \leq \alpha$. 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_α^λ, 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

$$\mathcal{S}_d^s(R) := \left\{ f \in L^1\big(\mathbb{R}^d\big) \cap L^2\big(\mathbb{R}^d\big) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\widehat{f}(\xi)|^2 \mathrm{d}\xi \leq (2\pi)^d R^2 \right\}, \tag{14}$$

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 $p - q \in \mathcal{S}_d^s(R)$ the probability of type II error of Δ_α^λ 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

$$(T_\lambda f)(x) := \int_{\mathbb{R}^d} f(x) \mathcal{K}_\lambda(x,y)\,\mathrm{d}y \tag{15}$$

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 $\mathcal{K}_\lambda := h_{k_\lambda,p}^{\mathrm{KSD}}$ 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 $h_{k_\lambda,p}^{\mathrm{KSD}}$ 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.

Theorem 1. *(i) Let $\sigma_{2,\lambda}^2 := \mathbb{E}\big[h_\lambda(Z,Z')^2\big]$. Assume $\|p\|_\infty \leq M$ and $\|q\|_\infty \leq M$ for some $M > 0$. For $\lambda \in (0,\infty)^d$ with $\lambda_1 \cdots \lambda_d < 1$, $\alpha \in (0, e^{-1})$, $\beta \in (0,1)$, $B_1 \geq \frac{2}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$, if*

$$\|p-q\|_2^2 \;\geq\; \|(p-q) - T_\lambda(p-q)\|_2^2 + C \frac{N}{L} \frac{\ln(1/\alpha)}{\beta} \sigma_{2,\lambda} \quad \textit{for some constant } C > 0,$$

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

(ii) Fix $R > 0$ and $s > 0$, and consider the bandwidths $\lambda_i^* := (N/L)^{2/(4s+d)}$ for $i = 1, \dots, d$. 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)

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

[4] Level is non-asymptotic for the goodness-of-fit case when using a parametric bootstrap (Schrab et al., 2022).

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. 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. 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 we recall that $L \leq N^2$ 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 $N < L < N^2$ 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 $L \leq N$ then there is no guarantee that the rate converges to 0.

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 (16). Then, using the incomplete U-statistic $\overline{U}_\lambda$, we define our aggregated test Δ_α^Λ as

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

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

$$u_\alpha \coloneqq \sup\nolimits_{B_3}\left\{u \in \left(0, \min_{\lambda\in\Lambda} w_\lambda^{-1}\right) : \frac{1}{B_2}\sum_{b=1}^{B_2} \mathbf{1}\left(\max_{\lambda\in\Lambda}\left(\widetilde{U}_\lambda^b - U_\lambda^{\bullet\lceil B_1(1-uw_\lambda)\rceil}\right) > 0\right) \leq \alpha\right\}, \quad (16)$$

where the wild bootstrapped incomplete U-statistics $\widetilde{U}_\lambda^1, \dots, \widetilde{U}_\lambda^{B_2}$ computed as in Equation (11) 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 (10), 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) Let $\sigma_{2,\lambda}^2 \coloneqq \mathbb{E}\big[h_\lambda(Z,Z')^2\big]$. Assume $\|p\|_\infty \leq M$ and $\|q\|_\infty \leq M$ for some $M > 0$. Consider a collection Λ such that $\lambda_1\cdots\lambda_d < 1$ for all $\lambda \in \Lambda$. For $\alpha \in (0, e^{-1})$, $B_1 \geq \frac{2}{\alpha^2}\left(\ln\left(\frac{8}{\beta}\right) + \alpha(1-\alpha)\right)$, $B_2 \geq \frac{8}{\alpha^2}\ln\left(\frac{2}{\beta}\right)$, $B_3 \geq \log_2\left(\frac{4}{\alpha}\min_{\lambda\in\Lambda} w_\lambda^{-1}\right)$, if*

$$\|p-q\|_2^2 \;\geq\; \min_{\lambda\in\Lambda}\left(\|(p-q) - T_\lambda(p-q)\|_2^2 + C\frac{N}{L}\frac{\ln(1/(\alpha w_\lambda))}{\beta}\sigma_{2,\lambda}\right) \text{ for some constant } C > 0,$$

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

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

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

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 $\left(\ln(\ln(N))/N\right)^{2s/(4s+d)}$ 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 (12), 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 (12) 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

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

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 G, uses the exponential concentration bound of Kim et al. (2022, Theorem 6.3) for permuted complete U-statistics. As discussed by Kim et al. (2022, Section 8.3), their proposed sample-splitting method can also be used to obtain the correct dependency on α.

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. Details on the experiments (e.g. model/test parameters) are presented in Appendix B.

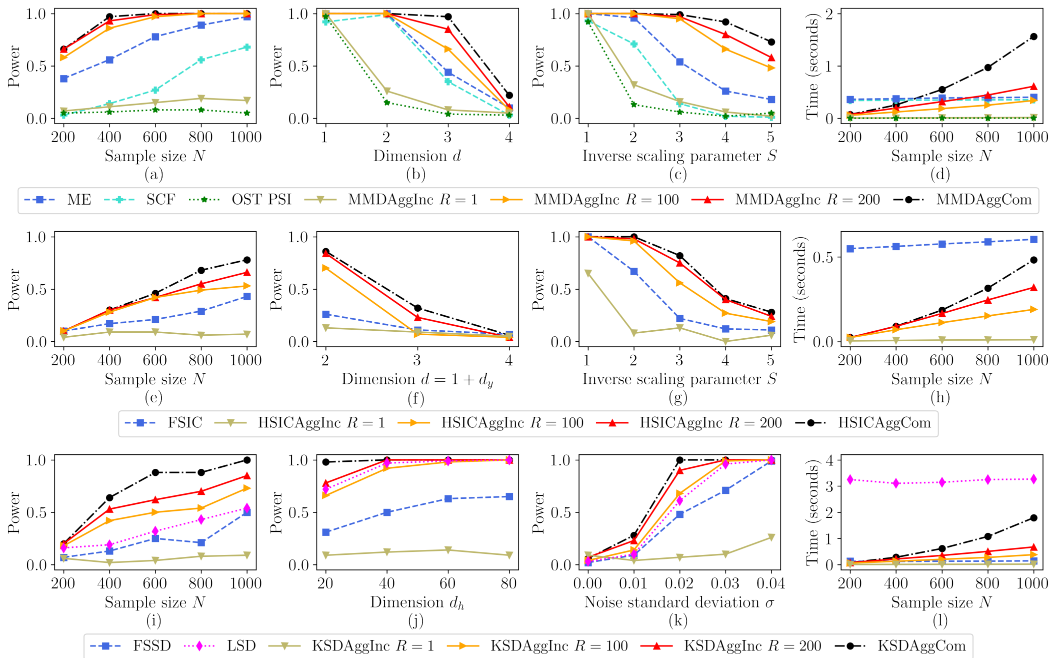

Figure 1: Two-sample *(a–d)* and independence *(e–h)* experiments using perturbed uniform densities. Goodness-of-fit *(i–l)* experiment using a Gaussian-Bernoulli Restricted Boltzmann Machine. The power results are averaged over 100 repetitions and the run times over 20 repetitions.

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, and LSD (Grathwohl et al., 2020) which uses a neural network to learn the Stein discrepancy (see Appendix B 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, we note that the run times 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, 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.

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.

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.

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.

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.

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

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.

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

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).

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://anonymous.4open.science/r/agginc-10EF/README.md`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (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 B.
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 B.
   (b) Did you mention the license of the assets? [Yes] See Appendix B.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] URL in Section 1: `https://anonymous.4open.science/r/agginc-10EF/README.md`.
   (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]