KSD Aggregated Goodness-of-fit Test

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

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 investigate properties of goodness-of-fit tests based on the Kernel Stein Discrepancy (KSD). We introduce a strategy to construct a test, called KSDAGG, which aggregates multiple tests with different kernels. KSDAGG avoids splitting the data to perform kernel selection (which leads to a loss in test power), and rather maximises the test power over a collection of kernels. We provide theoretical guarantees on the power of KSDAGG: we show it achieves the smallest uniform separation rate of the collection, up to a logarithmic term. For compactly supported densities with bounded score function for the model, we derive the rate for KSDAGG over restricted Sobolev balls; this rate corresponds to the minimax optimal rate over unrestricted Sobolev balls, up to an iterated logarithmic term. KSDAGG can be computed exactly in practice as it relies either on a parametric bootstrap or on a wild bootstrap to estimate the quantiles and the level corrections. In particular, for the crucial choice of bandwidth of a fixed kernel, it avoids resorting to arbitrary heuristics (such as median or standard deviation) or to data splitting. We find on both synthetic and real-world data that KSDAGG outperforms other state-of-the-art quadratic-time adaptive KSD-based goodness-of-fit testing procedures.

1 Introduction

Kernel selection remains a fundamental question in kernel-based nonparametric hypothesis testing, as it significantly impacts the test power. Kernel selection has attracted a significant interest in the literature, and a number of methods have been proposed in different settings, such as in the two-sample, independence and goodness-of-fit testing frameworks. Those methods include using heuristics (Gretton et al., 2012a), relying on data splitting (Gretton et al., 2012b; Sutherland et al., 2017; Kübler et al., 2022), learning deep kernels (Grathwohl et al., 2020; Liu et al., 2020), working in the post-selection inference framework (Yamada et al., 2019; Lim et al., 2019, 2020; Kübler et al., 2020; Freidling et al., 2021), to name but a few.

In this work, we focus on aggregated tests, which have been investigated for the two-sample problem by Fromont et al. (2013), Kim et al. (2022) and Schrab et al. (2021) using the Maximum Mean Discrepancy (MMD; Gretton et al., 2012a), and for the independence problem by Albert et al. (2022) and Kim et al. (2022) using the Hilbert Schmidt Independence Criterion (HSIC; Gretton et al., 2005). We extend the use of aggregated tests to the goodness-of-fit setting, where we are given a model and some samples, and test whether the samples have been drawn from the model. We employ the Kernel Stein Discrepancy (KSD; Chwialkowski et al., 2016; Liu et al., 2016) as our test statistic, which is

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

an ideal measure of distance for this setting: it admits an estimator which can be computed without requiring samples from the model, and does not require the model to be normalised. To the best of our knowledge, ours represents the first aggregation procedure for the KSD test in the literature.

Related work. Fromont et al. (2012, 2013) introduced non-asymptotic aggregated tests for the two-sample problem with the equal sample sizes following a Poisson process, using as test statistic an unscaled version of the MMD. They also provided theoretical results in terms of uniform separation rates using a wild bootstrap. Albert et al. (2022) then proposed an aggregated test for the independence problem using the HSIC, with guarantees using a theoretical quantile, but relying on permutations to obtain the test threshold in practice. Kim et al. (2022) then extended those results to also hold for the estimated quantile, and generalised the two-sample results to hold for the MMD estimator with different sample sizes using a wild bootstrap. All those aforementioned results were proved for the Gaussian kernel only. Schrab et al. (2021) generalised the two-sample results to hold for a wide range of kernels using either a wild bootstrap or a permutation-based procedure, and provided optimality results which hold with fewer restrictions on the class of functions. Our work builds and extends on the above results: we consider the goodness-of-fit framework, where we have samples from only one of the two densities. The main challenges arise from working with the Stein kernel that defines the KSD test statistic: for example, we lose the transition invariant property of the kernel which is crucial to work in the Fourier domain. We also point out the very relevant work of Balasubramanian et al. (2021) who considered adaptive MMD-based goodness-of-fit tests and studied their uniform separation rates over Sobolev balls in the asymptotic regime. More generally, Li and Yuan (2019) studied asymptotic adaptive kernel-based tests for the three testing frameworks. Tolstikhin et al. (2016) studied minimax rates for MMD estimators using radial universal kernel (Sriperumbudur et al., 2011). Finally, Schrab et al. (2022) extends this work, together with those of Albert et al. (2022) and Schrab et al. (2021), to construct efficient aggregated tests for goodness-of-fit, independence, and two-sample testing, using incomplete U-statistics. They quantify exactly the cost in the minimax rate over Sobolev balls incurred for computational efficiency. See Appendix B for other relevant works.

Contributions. We propose a solution to the fundamental kernel selection problem for the widely-used KSD goodness-of-fit tests: we construct an adaptive test KSDAGG which aggregates multiple tests with different kernels. Our contribution is in showing, both theoretically and experimentally, that the aggregation procedure works in this novel setting in which it has never been considered before. We consider the kernel selection framework; this general setting has many applications including the one of kernel bandwidth selection. Our aggregated test allows for two numerical methods for estimating the test thresholds: the wild bootstrap and the parametric bootstrap (a procedure unique to the goodness-of-fit framework). We conduct a theoretical analysis: we provide a lower bound on the uniform separation rate (Ingster, 1987, 1989, 1993a,b; Baraud, 2002) of KSDAGG, a condition which guarantees test power. We discuss the implementation of KSDAGG and experimentally validate our proposed approach on benchmark problems, not only on datasets classically used in the literature but also on original data obtained using state-of-the-art generative models (i.e. Normalizing Flows). We observe, both on synthetic and real-world data, that KSDAGG obtains higher power than other KSD-based adaptive state-of-the-art tests. Contributing to the real-world applications of these goodness-of-fit tests, we provide publicly available code to allow practitioners to employ our method: `https://github.com/antoninschrab/ksdagg-paper`.

2 Notation

We consider the goodness-of-fit problem where given access to a known model probability density p (which can be unnormalised since we only need its score function $\nabla \log p(\cdot)$) and to some i.i.d. d-dimensional samples $\mathbb{X}_N := (X_i)_{i=1}^N$ drawn from an unknown density q, we want to decide whether $p \neq q$ holds. This can be expressed as a statistical hypothesis testing problem with null hypothesis $\mathcal{H}_0 : p = q$ and alternative $\mathcal{H}_a : p \neq q$.

As a measure of distance between p and q, we use the *Kernel Stein Discrepancy (KSD)* introduced by Chwialkowski et al. (2016) and Liu et al. (2016). For a kernel k, the KSD is the Maximum Mean Discrepancy (MMD; Gretton et al., 2012a) between p and q using the Stein kernel associated to k

$$\begin{aligned}\mathrm{KSD}^2_{p,k}(q) := \mathrm{MMD}^2_{h_{p,k}}(p,q) &:= \mathbb{E}_{q,q}[h_{p,k}(X,Y)] - 2\mathbb{E}_{p,q}[h_{p,k}(X,Y)] + \mathbb{E}_{p,p}[h_{p,k}(X,Y)] \\ &= \mathbb{E}_{q,q}[h_{p,k}(X,Y)]\end{aligned}$$

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

$$h_{p,k}(x,y) \coloneqq \left(\nabla \log p(x)^\top \nabla \log p(y)\right) k(x,y) + \nabla \log p(y)^\top \nabla_x k(x,y) + \nabla \log p(x)^\top \nabla_y k(x,y) + \sum_{i=1}^{d} \frac{\partial}{\partial x_i \partial y_i} k(x,y)$$

and satisfies the *Stein identity* $\mathbb{E}_p[h_{p,k}(X,\cdot)] = 0$. Additional background details on the KSD are presented in Appendix B. A quadratic-time *KSD estimator* can be computed as the U-statistic (Hoeffding, 1992)

$$\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) \coloneqq \frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} h_{p,k}(X_i, X_j). \tag{1}$$

In this work, the model density p is always known, we do not always explicitly write the dependence of p for all variables (as we do for $\mathrm{KSD}^2_{p,k}$ and $h_{p,k}$). We assume that the kernel k is such that

$$\mathrm{KSD}^2_{p,k}(q) = \mathbb{E}_{q,q}[h_{p,k}(X,Y)] < \infty \qquad \text{and} \qquad C_k \coloneqq \mathbb{E}_{q,q}[h_{p,k}(X,Y)^2] < \infty. \tag{2}$$

We now address the requirements for consistency (*i.e.* test power converges to 1 as the sample size goes to ∞) of the Stein 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 $\mathbb{E}_q \left\| \nabla \left(\log \frac{p(X)}{q(X)} \right) \right\|_2^2 < \infty$.

We use the notations $\mathbb{P}_p$ and $\mathbb{P}_q$ to denote the probability under the model p and under q, respectively. Given a kernel $\kappa\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a function $f\colon \mathbb{R}^d \to \mathbb{R}$ in $L^2\left(\mathbb{R}^d\right)$, we consider the *integral transform* T_κ defined as

$$\left(T_\kappa f\right)(y) \coloneqq \int_{\mathbb{R}^d} \kappa(x,y) f(x) \,\mathrm{d}x$$

for $y \in \mathbb{R}^d$. When the kernel κ is translation invariant, the integral transform corresponds to a convolution. However, the lack of translation invariance of the Stein kernel introduces new challenging problems. First, working the integral transform of the Stein kernel is more complicated since the operation it does not correspond to a simple convolution. For the expectation of the Stein kernel squared, $C_k \coloneqq \mathbb{E}_{q,q}[h_{p,k}(X,Y)^2]$, it is not possible to extract the bandwidth parameter λ outside of the expectation as it is the case for the usual kernel k when using either the MMD or HSIC.

3 Construction of tests and bounds

We now introduce the single and aggregated KSD tests. We show that these control the probability of type I error as desired, and provide conditions for the control of the probability of type II error.

3.1 Single test

We first construct a KSD test for a fixed kernel k as proposed by Chwialkowski et al. (2016) and Liu et al. (2016). To estimate the test threshold, we can either use a wild bootstrap (Shao, 2010; Leucht and Neumann, 2013; Fromont et al., 2012; Chwialkowski et al., 2014) or a parametric bootstrap (Stute et al., 1993). Both methods work by simulating sampling values $\left(\bar{K}_k^1, \ldots, \bar{K}_k^{B_1}\right)$ from the (asymptotic) distribution of $\widehat{\mathrm{KSD}}^2_{p,k}$ under the null hypothesis and estimating the $(1-\alpha)$-quantile using a Monte Carlo approximation[1]

$$\widehat{q}^k_{1-\alpha} \coloneqq \inf\left\{ u \in \mathbb{R} : 1-\alpha \le \frac{1}{B_1} \sum_{b=1}^{B_1} \mathbb{1}\left(\bar{K}_k^b \le u\right) \right\} = \bar{K}_k^{\bullet \lceil B_1(1-\alpha) \rceil} \tag{3}$$

where $\bar{K}_k^{\bullet 1} \le \cdots \le \bar{K}_k^{\bullet B_1}$ are the sorted elements $\left(\bar{K}_k^1, \ldots, \bar{K}_k^{B_1}\right)$. The single test is then defined as

$$\Delta^k_\alpha(\mathbb{X}_N) \coloneqq \mathbb{1}\left(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) > \widehat{q}^k_{1-\alpha}\right).$$

[1] We do not write explicitly the dependence of $\widehat{q}^k_{1-\alpha}$ on other variables, but those are implicitly considered when writing probabilistic statements.

For the *parametric bootstrap*, we directly draw new samples $(X'_i)_{i=1}^N$ from the model p (it might not always be possible to so) and compute the KSD

$$\bar{K}_k \coloneqq \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} h_{p,k}(X'_i, X'_j). \tag{4}$$

For the *wild bootstrap*, we first generate n i.i.d. Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$ taking values in $\{-1, 1\}^N$, and then compute

$$\bar{K}_k \coloneqq \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} \epsilon_i \epsilon_j h_{p,k}(X_i, X_j). \tag{5}$$

Both these processes are then repeated B_1 times.

Since it uses samples from the model p, the parametric bootstrap (Stute et al., 1993) results in a test with non-asymptotic level α. This comes at the cost of being computationally more expensive and assuming that we are able to sample from p (which may be out of reach in some settings). Conversely, the wild bootstrap has the advantage of not requiring to sample from p, which makes it computationally more efficient as only one kernel matrix needs to be computed, but it only achieves the desired level α asymptotically (Shao, 2010; Leucht and Neumann, 2013; Leucht, 2012; Chwialkowski et al., 2014, 2016) assuming Lipschitz continuity of $h_{p,k}$ (see Appendix C for details). Note that we cannot obtain a non-asymptotic level for the wild bootstrap by relying on the result of Romano and Wolf (2005, Lemma 1) as done in the two-sample framework by Fromont et al. (2013) and Schrab et al. (2021). This is because in our case $\bar{K}_k$ and $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)$ are not exchangeable variables under the null hypothesis, due to the asymmetry of the KSD statistic with respect to p and q.

Having discussed control of the probability of type I error of the single test Δ^k_α, we now provide a condition on $\|p - q\|_2$ which ensures that the probability of type II error is controlled by some $\beta \in (0, 1)$. The smallest such value of $\|p - q\|_2$, provided that $p - q$ lies in some given class of functions, is called the *uniform separation rate* (Ingster, 1987, 1989, 1993a,b; Baraud, 2002).

Theorem 3.1. *Suppose the assumptions listed in Appendix A.1 hold, and let $\psi \coloneqq p - q$. There exists a positive constant C such that the condition*

$$\|\psi\|_2^2 \geq \left\|\psi - T_{h_{p,k}}\psi\right\|_2^2 + C \log\left(\frac{1}{\alpha}\right) \frac{\sqrt{C_k}}{\beta N}$$

guarantees control over the probability of type II error, such that $\mathbb{P}_q\big(\Delta^k_\alpha(\mathbb{X}_N) = 0\big) \leq \beta$.

Theorem 3.1, which is proved in Appendix H.1, provides a power guaranteeing condition consisting of two terms. The first term $\|\psi - T_{h_{p,k}}\psi\|_2^2$ indicates the size of the effect of the Stein operator on the difference in densities $\psi \coloneqq p - q$, and is a measure of distance from the null (where this quantity is zero). The second term $\log\big(1/\alpha\big)(\beta N)^{-1}\sqrt{C_k}$ is obtained from upper bounding the variance of the KSD U-statistic, it depends on the expectation of the squared Stein kernel $C_k \coloneqq \mathbb{E}_{q,q}[h_{p,k}(X, Y)^2]$. This second term also controls the quantile of the test.

3.2 Aggregated test

We can now introduce our aggregated test, which is motivated by the earlier works of Fromont et al. (2012, 2013), Albert et al. (2022), and Schrab et al. (2021) for different testing frameworks.

We compute $\widetilde{K}^1_k, \ldots, \widetilde{K}^{B_2}_k$ further simulated KSD values from the null hypothesis obtained using either a parametric bootstrap or a wild bootstrap as in Equations (4) or (5), respectively. Consider a finite collection of kernels $\mathcal{K}$ satisfying the properties presented in Section 2. We construct an aggregated test $\Delta^{\mathcal{K}}_\alpha$, called KSDAGG, which rejects the null hypothesis if one of the single tests $\left(\Delta^k_{u_\alpha w_k}\right)_{k \in \mathcal{K}}$ rejects it, that is

$$\Delta^{\mathcal{K}}_\alpha(\mathbb{X}_N) \coloneqq \mathbb{1}\big(\Delta^k_{u_\alpha w_k}(\mathbb{X}_N) = 1 \text{ for some } k \in \mathcal{K}\big)\,.$$

The levels of the single tests are adjusted to ensure the aggregated test has the prescribed level α. This adjustment is performed by introducing positive weights $(w_k)_{k \in \mathcal{K}}$ satisfying $\sum_{k \in \mathcal{K}} w_k \leq 1$ and some correction

$$u_\alpha \coloneqq \sup\left\{u \in \left(0, \min_{k \in \mathcal{K}} w_k^{-1}\right) : \widehat{P}_u \leq \alpha\right\} \tag{6}$$

Algorithm 1 KSDAGG

Inputs: samples $\mathbb{X}_N = (x_i)_{i=1}^N$, density p or score function $\nabla \log p(\cdot)$, finite kernel collection $\mathcal{K}$, weights $(w_k)_{k\in\mathcal{K}}$, level $\alpha \in (0, e^{-1})$, estimation parameters $B_1, B_2, B_3 \in \mathbb{N}$, parametric or wild bootstrap
Output: 0 (fail to reject $\mathcal{H}_0$) or 1 (reject $\mathcal{H}_0$)
Algorithm:
for $k \in \mathcal{K}$ **do**
  Compute $\left(\bar{K}_k^b\right)_{1\le b\le B_1}$ as in Equations (4) or (5)
  Sort in ascending order to obtain $\left(\bar{K}_k^{\bullet b}\right)_{1\le b\le B_1}$
  Compute $\left(\widetilde{K}_k^b\right)_{1\le b\le B_2}$ as in Equations (4) or (5)
$u_{\min} = 0$, $u_{\max} = \min_{k\in\mathcal{K}} w_k^{-1}$
for $t = 1, \dots, B_3$ **do**
  $u = \frac{1}{2}(u_{\min} + u_{\max})$, $\widehat{P}_u = \dfrac{1}{B_2}\displaystyle\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{k\in\mathcal{K}}\left(\widetilde{K}_k^b - \bar{K}_k^{\bullet\lceil B_1(1-uw_k)\rceil}\right) > 0\right)$
  if $\widehat{P}_u \le \alpha$ **then** $u_{\min} = u$ **else** $u_{\max} = u$
$u_\alpha = u_{\min}$
if $\max_{k\in\mathcal{K}}\left(\widehat{\mathrm{KSD}}_{p,k}^2(\mathbb{X}_N) - \bar{K}_k^{\bullet\lceil B_1(1-u_\alpha w_k)\rceil}\right) > 0$ **then return** 1 **else return** 0
Time complexity: $\mathcal{O}\left(|\mathcal{K}|(B_1 + B_2)\, N^2\right)$

where

$$\widehat{P}_u \coloneqq \frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{k\in\mathcal{K}}\left(\widetilde{K}_k^b - \bar{K}_k^{\bullet\lceil B_1(1-uw_k)\rceil}\right) > 0\right)$$

is a Monte Carlo approximation of the probability of type I error of our aggregated test with correction u

$$P_u \coloneqq \mathbb{P}_p\left(\max_{k\in\mathcal{K}}\left(\widehat{\mathrm{KSD}}_{p,k}^2(\mathbb{X}_N) - \widehat{q}_{1-uw_k}^k\right) > 0\right).$$

To compute u_α, we estimate the supremum in Equation (6) by performing B_3 steps of the bisection method. Detailed pseudocode for KSDAGG is provided in Algorithm 1, and details regarding our aggregation procedure are provided in Appendix F. We discuss potential limitations of KSDAgg in Appendix G.

We verify in the next proposition that performing this correction indeed ensures that our aggregated test $\Delta_\alpha^{\mathcal{K}}$ has the prescribed level α.

Proposition 3.2. *For $\alpha \in (0, 1)$ and a collection of kernels $\mathcal{K}$, the aggregated test $\Delta_\alpha^{\mathcal{K}}$ satisfies*

$$\mathbb{P}_p\left(\Delta_\alpha^{\mathcal{K}}(\mathbb{X}_N) = 1\right) \le \alpha$$

asymptotically using a wild bootstrap (with Lipschitz continuity of $h_{p,k}$ required) and non-asymptotically using a parametric bootstrap.

The proof of Proposition 3.2 is presented in Appendix H.2. Details about the wild bootstrap case are presented in Appendix C. We now provide guarantees for the power of our aggregated test KSDAGG in terms of its uniform separation rate.

Theorem 3.3. *Suppose the assumptions listed in Appendix A.2 hold, and let $\psi \coloneqq p - q$. There exists a positive constant C such that if*

$$\|\psi\|_2^2 \ge \min_{k\in\mathcal{K}}\left(\left\|\psi - T_{h_{p,k}}\psi\right\|_2^2 + C\log\left(\frac{1}{\alpha w_k}\right)\frac{\sqrt{C_k}}{\beta N}\right)$$

then the probability of type II error of $\Delta_\alpha^{\mathcal{K}}$ is controlled by β, that is, $\mathbb{P}_q\left(\Delta_\alpha^{\mathcal{K}}(\mathbb{X}_N) = 0\right) \le \beta$.

The proof Theorem 3.3 can be found in Appendix H.3, it relies on upper bounding a probability of intersections of events by the minimum of the probabilities of each event, on checking that the assumptions of Theorem 3.1 are satisfied for the adjusted levels, and on applying Theorem 3.1. We observe that the aggregation procedure allows to achieve the smallest uniform separation rate of the single tests $\left(\Delta_\alpha^k\right)_{k\in\mathcal{K}}$ up to some logarithmic weighting term $\log(1/w_k)$.

3.3 Bandwidth selection

A specific application of the setting we have considered is the problem of bandwidth selection for a fixed kernel. Given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the function

$$k_\lambda(x, y) \coloneqq k\Big(\frac{x}{\lambda}, \frac{y}{\lambda}\Big)$$

is also a kernel for any bandwidth $\lambda > 0$. A common example is the Gaussian kernel, for which we have $k(x, y) = \exp(-\|x - y\|_2^2)$ and $k_\lambda(x, y) = \exp\big(-\|x - y\|_2^2/\lambda^2\big)$. As shown by Gorham and Mackey (2017), a more appropriate kernel for goodness-of-fit testing using the KSD is the IMQ (inverse multiquadric) kernel, which is defined with $k(x, y) = \Big(1 + \|x - y\|_2^2\Big)^{-\beta_k}$ for a fixed parameter $\beta_k \in (0, 1)$ as

$$k_\lambda(x, y) = \left(1 + \frac{\|x - y\|_2^2}{\lambda^2}\right)^{-\beta_k} = \lambda^{2\beta_k}\Big(\lambda^2 + \|x - y\|_2^2\Big)^{-\beta_k} \propto \Big(\lambda^2 + \|x - y\|_2^2\Big)^{-\beta_k} \tag{7}$$

which is the well-known form of the IMQ kernel with parameters $\lambda > 0$ and $\beta_k \in (0, 1)$. Note that it is justified to consider the kernel up to a multiplicative constant because our single and aggregated tests are invariant under this kernel transformation.

In practice, as suggested by Gretton et al. (2012a), the bandwidth is often set to a heuristic such as the median or the standard deviation of the L^2-distances between the samples $(X_i)_{i=1}^N$, however, these are arbitrary choices with no theoretical guarantees. Another common approach proposed by Gretton et al. (2012b) for the linear-time setting, and extended to the quadratic-time setting by Liu et al. (2020), is to resort to data splitting in order to select a bandwidth on held-out data, by maximising for a proxy for asymptotic power (see Section 4.1 for details). Both methods were originally proposed for the two-sample problem, but extend straightforwardly to the goodness-of-fit setting.

By considering a kernel collection $\mathcal{K}_\Lambda = \{k_\lambda : \lambda \in \Lambda\}$ for a collection of bandwidths Λ, we can use our aggregated test KSDAGG to test multiple bandwidths using all the data and without resorting to arbitrary heuristics. We now obtain an expression for the uniform separation rate of $\Delta_\alpha^{\mathcal{K}_\Lambda}$ in terms of the bandwidths $\lambda \in \Lambda$.

Corollary 3.4. *Suppose the assumptions listed in Appendix A.2 hold for $\mathcal{K}_\Lambda = \{k_\lambda : \lambda \in \Lambda\}$. Let $\psi \coloneqq p - q$ and consider C_{k_λ} as defined in Equation* (2)*. There exists $C > 0$ such that the condition*

$$\|\psi\|_2^2 \geq \min_{\lambda \in \Lambda}\left(\Big\|\psi - T_{h_{p,k_\lambda}}\psi\Big\|_2^2 + C\log\left(\frac{1}{\alpha w_\lambda}\right)\frac{\sqrt{C_{k_\lambda}}}{\beta N}\right)$$

ensures control over the probability of type II error of the aggregated test $\mathbb{P}_q\big(\Delta_\alpha^{\mathcal{K}_\Lambda}(\mathbb{X}_N) = 0\big) \leq \beta$.

Corollary 3.4 follows from applying Theorem 3.3 to the collection of kernels $\mathcal{K}_\Lambda$. Our results hold with great generality as we have not imposed any restrictions on $\psi \coloneqq p - q$ such as assuming it belongs to a specific regularity class. For this reason, the dependence on λ in the terms $\|\psi - T_{h_{p,k_\lambda}}\psi\|_2^2$ and $\log\big(1/(\alpha w_\lambda)\big)(\beta N)^{-1}\sqrt{C_{k_\lambda}}$ is not explicit. In Section 3.4, we characterise this dependence with regularity assumptions on $\psi \coloneqq p - q$, and derive uniform seperation rates in terms of the sample size.

3.4 Uniform separation rates over restricted Sobolev balls

In this section, we derive uniform separation rates with stronger assumptions on p and q. This provides settings in which the power guaranteeing conditions of Theorems 3.1 and 3.3 are satisfied, and illustrates the interactions between the two terms in those conditions. We make the following assumptions.

- The model density p is strictly positive on its connected and compact support $S \subseteq \mathbb{R}^d$.
- The score function $\nabla \log p(x)$ is continuous and bounded on S.
- The support of the density q is a connected and compact subset of S.
- The kernel used is a scaled Gaussian kernel $k_\lambda(x, y) \coloneqq \lambda^{2-d}\exp\Big(-\|x - y\|_2^2 / \lambda^2\Big)$.

In particular, any strictly positive twice-differentiable density truncated to some d-dimensional interval will satisfy the assumptions for the model p. Any density truncated to the same d-dimensional interval will satisfy the assumption for the density q. As an example, one can consider truncated normal densities.

For smoothness parameter $s > 0$, radius $R > 0$ and dimension $d \in \mathbb{N} \setminus \{0\}$, the Sobolev ball $\mathcal{S}_d^s(R)$ is defined as the function space

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

where $\widehat{f}$ denotes the Fourier transform of f. For s, t, d, R, L all strictly positive, we define the restricted Sobolev ball $\mathcal{S}_d^{s,t}(R, L)$ as containing all functions $f \in \mathcal{S}_d^s(R)$ satisfying

$$\int_{\|\xi\|_2 \leq t} \left| \widehat{f}(\xi) \right|^2 \mathrm{d}\xi \leq \frac{1}{L} \int_{\mathbb{R}^d} \left| \widehat{f}(\xi) \right|^2 \mathrm{d}\xi. \tag{8}$$

With the Sobolev assumption $p - q \in \mathcal{S}_d^s(R)$, the densities p and q can differ at any frequencies. The restricted Sobolev assumption $p - q \in \mathcal{S}_d^{s,t}(R, L)$ does not include the case in which the densities p and q differ only at low frequencies due to the additional restriction in Equation (8).

Theorem 3.5. *Suppose the assumptions listed in Appendix A.3 hold for $\mathcal{K}_\Lambda = \{k_\lambda : \lambda \in \Lambda\}$. Let $\psi := p - q$. We show that there exists some $L > 0$ such that if $p - q \in \mathcal{S}_d^{s,t}(R, L)$ then (i) & (ii) hold.*

(i) Under the assumptions of Theorem 3.1, for the KSD test with bandwidth $\lambda := N^{-2/(4s+d)}$, the condition

$$\|\psi\|_2^2 \geq C N^{-4s/(4s+d)}.$$

guarantees control over the probability of type II error, so that $\mathbb{P}_q\big(\Delta_\alpha^\lambda(\mathbb{X}_N) = 0\big) \leq \beta$.

(ii) Under the assumptions of Theorem 3.3, for KSDAGG *with the collection*

$$\Lambda := \left\{ 2^{-\ell} : \ell \in \left\{ 1, \ldots, \left\lceil \frac{2}{d} \log_2\left(\frac{N}{\ln(\ln(N))} \right) \right\rceil \right\} \right\}$$

and weights $w_\lambda := 6/\pi^2 \ell^2$, we have $\mathbb{P}_q\big(\Delta_\alpha^\Lambda(\mathbb{X}_N) = 0\big) \leq \beta$ provided that

$$\|\psi\|_2^2 \geq C \left(\frac{N}{\ln(\ln(N))} \right)^{-4s/(4s+d)}.$$

The proof of Theorem 3.5 is presented in Appendix H.4. The uniform separation rate $N^{-4s/(4s+d)}$ in Theorem 3.5 is known to be optimal in the *minimax* sense over (unrestricted) Sobolev balls $\mathcal{S}_d^s(R)$ (Li and Yuan, 2019; Balasubramanian et al., 2021; Albert et al., 2022; Schrab et al., 2021). However, this rate is for the KSD test with bandwidth depending on the smoothness parameter $s > 0$ which is unknown in practice. Note that even methods which split the data to perform kernel selection are not able to select this bandwidth depending on s and cannot achieve the minimax rate. The aggregation procedure is adaptive to this unknown parameter s: the collection of bandwidths Λ does not depend on s (it can be implemented in practice). The price to pay for this adaptivity is only an iterated logarithmic factor in the minimax rate over Sobolev balls $\mathcal{S}_d^s(R)$. The uniform separation rates presented in Theorem 3.5 are over the restricted Sobolev balls $\mathcal{S}_d^{s,t}(R, L)$. Whether those rates can also be derived under less restrictive assumptions and in the more general setting of (unrestricted) Sobolev balls $\mathcal{S}_d^s(R)$ is a challenging problem, which is left for future work.

4 Implementation and experiments

We consider three different experiments based on a Gamma one-dimensional distribution, a Gaussian-Bernoulli Restricted Boltzmann Machine, and a Normalizing Flow for the MNIST dataset. We compare our proposed aggregated test KSDAGG against three alternatives: the KSD test which uses the median bandwidth, a test which splits the data to select an 'optimal' bandwidth according to a proxy for asymptotic test power, and a test which uses extra data for bandwidth selection. The 'extra data' test is designed simply to provide a best case oracle for the bandwidth selection procedure which maximises asymptotic test power, it cannot be used in practice (i.e. any extra samples from q would normally be incorporated into the sample being tested). In order to ensure that our tests always have correct levels for all bandwidth values, dimensions and sample sizes, we use the parametric bootstrap in our experiments.

4.1 Alternative bandwidth selection approaches

Gretton et al. (2012a) proposed to use the median heuristic as kernel bandwidth, it consists in the median of the L^2-distances between the samples given by

$$\lambda_{\text{med}} := \text{median}\{\|x_i - x_j\|_2 : 0 \leq i < j \leq N\}.$$

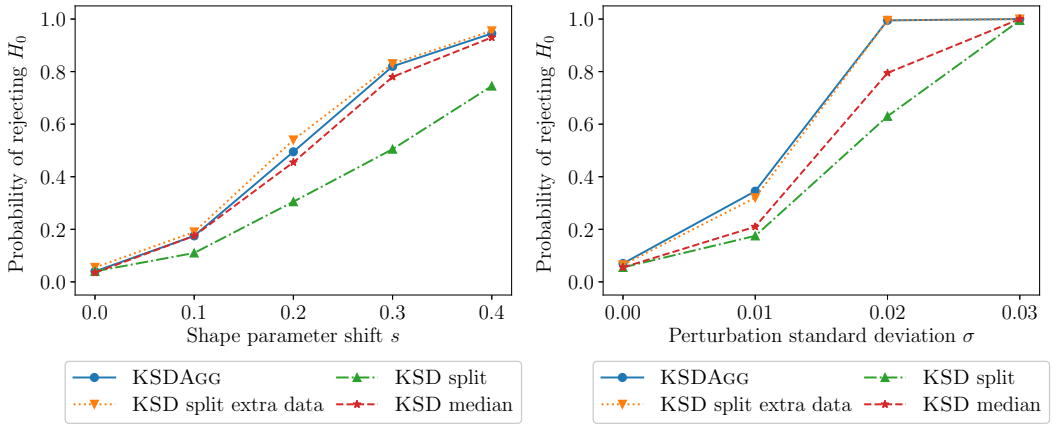
Gretton et al. (2012b) first proposed, for the two-sample problem using a linear-time MMD estimator, to split the data and to use half of it to select an 'optimal' bandwidth which maximises a proxy for asymptotic power. This procedure was extended to quadratic-time estimators and to the goodness-of-fit framework by Jitkrittum et al. (2017), Sutherland et al. (2017) and Liu et al. (2020). These strategies rely on the asymptotic normality of the test statistic under $\mathcal{H}_a$. In our setting, the asymptotic power proxy to maximise is the ratio

$$\widehat{\text{KSD}}^2_{p,k}(\mathbb{X}_N) \,/\, \widehat{\sigma}_{\mathcal{H}_a} \tag{9}$$

where $\widehat{\sigma}^2_{\mathcal{H}_a}$ is a closed-form regularised positive estimator of the asymptotic variance of $\widehat{\text{KSD}}^2_{p,k}$ under $\mathcal{H}_a$ (Liu et al., 2020, Equation 5). In our experiments, we also consider a test which has access to N extra samples drawn from q to select an 'optimal' bandwidth to run the KSD test on the original N samples $\mathbb{X}_N$. This test is interesting to compare to because it uses an 'optimal' bandwidth without being detrimental to power.

4.2 Experimental details

In our experiments, we use collections of bandwidths of the form $\Lambda(\ell_-, \ell_+) := \left\{2^i \lambda_{\text{med}} : i = \ell_-, \ldots, \ell_+\right\}$ for the median bandwidth λ_{med} and integers $\ell_- < \ell_+$ with uniform weights $w_\lambda := 1/(\ell_+ - \ell_- + 1)$. For the tests which split the data, we select the bandwidth out of the collection $\Lambda(\ell_-, \ell_+)$ which maximises the power proxy discussed in Section 4.1. All our experiments are run with level $\alpha = 0.05$ using the IMQ kernel defined in Equation (7) with parameter $\beta_k = 0.5$. We use a parametric bootstrap with $B_1 = B_2 = 500$ simulated KSD values to compute the adjusted test thresholds, and $B_3 = 50$ steps of bisection method to estimate the correction u_α in Equation (6). To estimate the probability of rejecting the null hypothesis, we average the test outputs across 200 repetitions. All experiments have been run on an AMD Ryzen Threadripper 3960X 24 Cores 128Gb RAM CPU at 3.8GHz, the runtime is of the order of a couple of hours (significant speedup can be obtained by using parallel computing). We have used the implementation of Jitkrittum et al. (2017) to sample from a Gaussian-Bernoulli Restricted Boltzmann Machine, and Phillip Lippe's implementation of MNIST Normalizing Flows, both under the MIT license.



(a) (b)

Figure 1: (a) Gamma distribution experiment. (b) Gaussian-Bernoulli Restricted Boltzmann Machine experiment.

4.3 Gamma distribution

For our first experiment, we consider a one-dimensional Gamma distribution with shape parameter 5 and scale parameter 5 as the model p. We draw 500 samples from a Gamma distribution with the same scale parameter 5 and with a shifted shape parameter $5 + s$ for $s \in \{0, 0.1, 0.2, 0.3, 0.4\}$. We consider the collection of bandwidths $\Lambda(0, 10)$.

The results we obtained are presented in Figure 1a. We observe that all tests have the prescribed level 0.05 under the null hypothesis, which corresponds to the case $s = 0$. As the shift parameter s increases, the two densities p and q become more different and rejection of the null becomes an easier task, thus the test power increases. Our aggregated test KSDAGG achieves the same power as the 'best case' bound on the performance of the asymptotic power heuristic, yielded by the splitting test with extra data. The median test obtains only slightly lower power, this closeness in power can be explained by the fact that this one-dimensional problem is a simple one. We note that the normal splitting test has significantly lower power: this is because, even though it uses an 'optimal' bandwidth, it is then run on only half the data, which results in a loss of power.

4.4 Gaussian-Bernoulli Restricted Boltzmann Machine

As first considered by Liu et al. (2016) for goodness-of-fit testing using the KSD, we consider a Gaussian-Bernoulli Restricted Boltzmann Machine. It is a graphical model with a binary hidden variable $h \in \{-1, 1\}^{d_h}$ and a continuous observable variable $x \in \mathbb{R}^d$. Those variables have joint density

$$p(x, h) = \frac{1}{Z} \exp\left(\frac{1}{2} x^\top B h + b^\top x + c^\top h - \frac{1}{2}\|x\|_2^2\right)$$

where Z is an unknown normalizing constant. By marginalising over h, we obtain the density p of x

$$p(x) = \sum_{h \in \{-1,1\}^{d_h}} p(x, h).$$

We can sample from it using a Gibbs sampler with 2000 burn-in iterations. We use the dimensions $d = 50$ and $d_h = 40$ as considered by Jitkrittum et al. (2017) and Grathwohl et al. (2020). Even though computing p is intractable for large dimension d_h, the score function admits a convenient closed form

$$\nabla \log p(x) = b - x + B \frac{\exp\big(2(B^\top x + c)\big) - 1}{\exp(2(B^\top x + c)) + 1}.$$

We draw the components of b and c from Gaussian standard distributions and sample Rademacher variables taking values in $\{-1, 1\}$ for the elements of B for the model p. We draw 1000 samples from a distribution q which is constructed in a similar way as p but with the difference that some Gaussian noise $\mathcal{N}(0, \sigma)$ is injected into the elements of B. We consider the standard deviations of the perturbations $\sigma \in \{0, 0.01, 0.02, 0.03\}$. We run our experiments with the collection of bandwidths $\Lambda(-20, 0)$ and provide the results in Figure 1b.

Again, we observe that our aggregated test KSDAGG matches the power obtained by the test which uses extra data to select an 'optimal' bandwidth. This means that, in this experiment, KSDAGG obtains the same power as the 'best' single test. The difference between KSDAGG and the median heuristic test is significant on this experiment, and the splitting test obtains lowest power of the four tests. Again, all tests have well-calibrated levels ($\sigma = 0$) and increasing the noise level σ results in more power for all the tests.

(a) (b) (c)

Figure 2: (a) Digits from the MNIST dataset. (b, c) Digits sampled from the Normalizing Flow.

4.5 MNIST Normalizing Flow

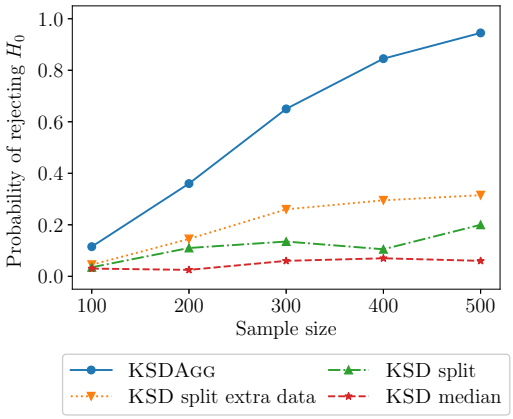
Finally, we consider a high-dimensional problem working with images in dimensions $28^2 = 784$. We consider a multi-scale Normalizing Flow (Dinh et al., 2017; Kingma and Dhariwal, 2018) which has been trained on the MNIST dataset (LeCun et al., 1998, 2010), it is a generative model which has a probability density p. As observed in Figure 2, some samples produced by the model can look exactly like MNIST digits, while other do not resemble digits. This Normalizing Flow has been trained to 'ideally' produce samples from the MNIST dataset. We are interested in whether or not we can detect the difference in densities. Given some images of digits, are we able to tell if those were sampled from the Normalizing Flow model? We consider the case where the images from q are sampled from the model (level experiment, confirming performance under $\mathcal{H}_0$), and the case where the samples from q are drawn from the true MNIST dataset (power experiment). The experiments are run with the collection of bandwidths $\Lambda(-20, 0)$. The results are displayed in Figure 3a and Figure 3b.

In Figure 3a, we observe that the four tests have correct levels (around 0.05) for the five different sample sizes considered (the small fluctuations about the designed test level can be explained by the fact that we are averaging 200 test outputs to estimate these levels). The well-calibrated levels obtained in Figure 3a demonstrate the validity of the power results presented in Figure 3b.

In Figure 3b, we observe that only our aggregated test KSDAGG obtains high power, that is, is able to detect that MNIST samples are not drawn from the Normalizing Flow. The power of the other tests increases only marginally as the sample size increases. We notice that the test which uses extra data to select an 'optimal' bandwidth performs poorly when compared to KSDAGG. This could be explained by the fact that this test selects the bandwidth using a proxy for the asymptotic power, and that in this high-dimensional setting, the asymptotic regime is not attained with sample sizes below 500. See Appendices D and E for details regarding 'optimal' bandwidths and for runtimes.

SAMPLE SIZE	KSDAGG	KSD MEDIAN	KSD SPLIT	KSD SPLIT EXTRA DATA
100	0.05	0.04	0.05	0.04
200	0.06	0.055	0.05	0.055
300	0.045	0.085	0.05	0.085
400	0.04	0.055	0.03	0.055
500	0.035	0.045	0.04	0.045

(a)



(b)

Figure 3: Normalizing Flow MNIST. (a) Level experiment. (b) Power experiment.

5 Acknowledgements

Antonin Schrab acknowledges support from the U.K. Research and Innovation (EP/S021566/1). 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.

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.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., et al. (2021). Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*.

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).

Barp, A., Briol, F., Duncan, A. B., Girolami, M. A., and Mackey, L. W. (2019). Minimum Stein discrepancy estimators. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems*.

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.

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.

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*.

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.

Fernandez, T. and Gretton, A. (2019). A maximum-mean-discrepancy goodness-of-fit test for censored data. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2966–2975. PMLR.

Fernández, T. and Rivera, N. (2022). A general framework for the analysis of kernel-based tests. *arXiv preprint arXiv:2209.00124*.

Fernandez, T., Rivera, N., Xu, W., and Gretton, A. (2020). Kernelized Stein discrepancy tests of goodness-of-fit for time-to-event data. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3112–3122. PMLR.

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.

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

Gorham, J. and Mackey, L. W. (2015). Measuring sample quality with Stein's method. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 226–234.

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

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.

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.

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

Ingster, Y. I. (1989). An asymptotically minimax test of the hypothesis of independence. *Journal of Soviet Mathematics*, 1(44:466–476).

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

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

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In Peters, J. and Sontag, D., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. PMLR.

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

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. (2019). A kernel Stein test for comparing latent variable models. *arXiv:1907.00586 [cs, stat]*. arXiv: 1907.00586.

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

Korba, A., Aubin-Frankowski, P., Majewski, S., and Ablin, P. (2021). Kernel Stein discrepancy descent. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 5719–5730. PMLR.

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., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

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

Leucht, A. (2012). Degenerate U- and V-statistics under weak dependence: Asymptotic theory and bootstrap consistency. *Bernoulli*, 18(2):552 – 585.

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, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*.

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.

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. (2019). Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141 – 1159.

Oates, C. J. and Girolami, M. A. (2016). Control functionals for quasi-Monte Carlo integration. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 56–65. JMLR.org.

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

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022). Efficient aggregated kernel tests using incomplete U-statistics. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

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

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).

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 6, pages 583–603. University of California Press.

Stein, C., Diaconis, P., Holmes, S., and Reinert, G. (2004). Use of exchangeable pairs in the analysis of simulations. *Lecture Notes-Monograph Series*, pages 1–26.

Stute, W., Manteiga, W. G., and Quindimil, M. P. (1993). Bootstrap based goodness-of-fit-tests. *Metrika*, 40(1):243–256.

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*.

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

Weckbecker, M., Xu, W., and Reinert, G. (2022). On RKHS choices for assessing graph generators via kernel Stein statistics. *arXiv preprint arXiv:2210.05746*.

Wenliang, L. K. and Kanagawa, H. (2020). Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*.

Xu, W. (2021). Generalised kernel Stein discrepancy (GKSD): A unifying approach for non-parametric goodness-of-fit testing. *arXiv preprint arXiv:2106.12105*.

Xu, W. (2022). Standardisation-function kernel Stein discrepancy: A unifying view on kernel Stein discrepancy tests for goodness-of-fit. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1575–1597. PMLR.

Xu, W. and Matsuda, T. (2020). A Stein goodness-of-fit test for directional distributions. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 320–330. PMLR.

Xu, W. and Matsuda, T. (2021). Interpretable Stein goodness-of-fit tests on Riemannian manifold. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11502–11513. PMLR.

Xu, W. and Reinert, G. (2021). A Stein goodness-of-test for exponential random graph models. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 415–423. PMLR.

Xu, W. and Reinert, G. (2022a). AgraSSt: Approximate graph Stein statistics for interpretable assessment of implicit graph generators. *arXiv preprint arXiv:2203.03673*.

Xu, W. and Reinert, G. (2022b). A kernelised Stein statistic for assessing implicit generative models. *arXiv preprint arXiv:2206.00149*.

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*.

Zhang, M., Key, O., Hayes, P., Barber, D., Paige, B., and Briol, F.-X. (2022). Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*.

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 Appendix G.
   (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 Appendices.
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] See link in Section 1: `https://github.com/antoninschrab/ksdagg-paper`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] 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 Section 4.2.
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 Section 4.2.
   (b) Did you mention the license of the assets? [Yes] See Section 4.2.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See link in Section 1: `https://github.com/antoninschrab/ksdagg-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]

Supplementary material for 'KSD Aggregated Goodness-of-fit Test'

A Assumptions

A.1 Assumptions for Theorem 3.1

- $\|q\|_\infty \leq M$ for some $M > 0$
- $\alpha \in (0, e^{-1})$
- $\beta \in (0, 1)$
- $B_1 \geq 3\big(\log\big(8/\beta\big) + \alpha(1-\alpha)\big)/\alpha^2$
- $C_k \geq 1$, where $C_k \coloneqq \mathbb{E}_{q,q}[h_{p,k}(X,Y)^2]$ as defined in Equation (2)
- assumption for parametric bootstrap: $n/\sqrt{C_k} \geq \ln(1/\alpha)$
- assumption for wild bootstrap: Lipschitz continuity of $h_{p,k}$
- constant C depends on M and d

A.2 Assumptions for Theorem 3.3 and Corollary 3.4

- $\|q\|_\infty \leq M$ for some $M > 0$
- $\alpha \in (0, e^{-1})$
- $\beta \in (0, 1)$
- $B_1 \geq 12\Big(\max_{k\in\mathcal{K}} w_k^{-2}\Big)\big(\log(8/\beta) + \alpha(1-\alpha)\big)/\alpha^2$
- $B_2 \geq 8\log\big(2/\beta\big)/\alpha^2$
- $B_3 \geq \log_2\big(4\min_{k\in\mathcal{K}} w_k^{-1}/\alpha\big)$
- $C_k \geq 1$ for $k \in \mathcal{K}$, where $C_k \coloneqq \mathbb{E}_{q,q}[h_{p,k}(X,Y)^2]$ as defined in Equation (2)
- positive weights $(w_k)_{k\in\mathcal{K}}$ satisfying $\sum_{k\in\mathcal{K}} w_k \leq 1$
- assumption for parametric bootstrap: $n/\sqrt{C_k} \geq \ln(1/\alpha w_k)$ for $k \in \mathcal{K}$
- assumption for wild bootstrap: Lipschitz continuity of $h_{p,k}$
- constant C depends on M and d

A.3 Assumptions for Theorem 3.5

- $\|q\|_\infty \leq M$ for some $M > 0$
- $\alpha \in (0, e^{-1})$
- $\beta \in (0, 1)$
- $B_1 \geq 12\Big(\max_{\lambda\in\Lambda} w_{k_\lambda}^{-2}\Big)\big(\log(8/\beta) + \alpha(1-\alpha)\big)/\alpha^2$
- $B_2 \geq 8\log\big(2/\beta\big)/\alpha^2$
- $B_3 \geq \log_2\big(4\min_{\lambda\in\Lambda} w_{k_\lambda}^{-1}/\alpha\big)$
- $C_{k_\lambda} \geq 1$ for $\lambda \in \Lambda$, where $C_{k_\lambda} \coloneqq \mathbb{E}_{q,q}[h_{p,k_\lambda}(X,Y)^2]$ as defined in Equation (2)
- assumption for parametric bootstrap: $n/\sqrt{C_{k_\lambda}} \geq \ln(1/\alpha w_{k_\lambda})$ for $\lambda \in \Lambda$
- assumption for wild bootstrap: Lipschitz continuity of h_{p,k_λ}
- constant C is independent of the sample size N

B Background on Kernel Stein Discrepancy

Background. Stein's methods (Stein, 1972) have been widely used in the machine learning and statistics communities. At the heart of this field is the fact that smooth densities p and q are equal if and only if

$$\mathbb{E}_q\big[f(x)^\top \nabla \log p(x) + \mathrm{Tr}\big(\nabla f(x)\big)\big] = 0$$

for smooth functions f vanishing at the boundaries. This is known as Stein's identity (Stein, 1972; Stein et al., 2004), which can be utilised by Gorham and Mackey (2015) for defining Stein discrepancies as

$$\sup_{f \in \mathcal{F}} \mathbb{E}_q\big[f(x)^\top \nabla \log p(x) + \mathrm{Tr}\big(\nabla f(x)\big)\big]$$

for some space of smooth functions $\mathcal{F}$ satisfying Stein's identity. As proposed by Oates and Girolami (2016), choosing $\mathcal{F}$ to be the unit ball in a Reproducing Kernel Hilbert Space (Aronszajn, 1950, RKHS) denoted $\mathcal{H}$, that is $\mathcal{F} := \{f : \|f\|_{\mathcal{H}} \leq 1\}$, we obtain the Kernel Stein discrepancy (KSD; Liu et al., 2016; Chwialkowski et al., 2016). The KSD can also be expressed in terms of the Stein kernel as presented in Section 2.

Related work. Stein methods have been used to construct goodness-of-fit tests in various types of data: directional data (Xu and Matsuda, 2020), time-to-event Data (Fernandez et al., 2020), data on Riemannian manifolds (Xu and Matsuda, 2021), conditional data (Jitkrittum et al., 2020) , graph data (Weckbecker et al., 2022; Xu and Reinert, 2021, 2022a), and generative data (Xu and Reinert, 2022b). We also point out the works of Fernandez and Gretton (2019) on an MMD-based goodness-of-fit test for censored data, of Key et al. (2021) on composite goodness-of-fit tests using either the MMD or KSD, of Jitkrittum et al. (2017) on a linear-time KSD goodness-of-fit test, of Gorham and Mackey (2017) on measuring sample quality with kernels and on KSD dominating weak convergence for some kernels, of Huggins and Mackey (2018) on random feature Stein discrepancies, of Korba et al. (2021) on the KSD Wasserstein gradient flow, of Oates et al. (2019) on convergence rates for a class of estimators based on Stein's method, and of Barp et al. (2019) on minimum Stein discrepancy estimators. Lim et al. (2019) and Kanagawa et al. (2019) use KSD tests for multiple model comparison and for comparing latent variable models, respectively. It is worth noting those score-based methods suffer from blindness to isolated components and mixing proportions (Wenliang and Kanagawa, 2020; Zhang et al., 2022). Xu (2021, 2022) present unified approaches for goodness-of-fit testing, and Fernández and Rivera (2022) introduces a general framework for analysing kernel-based tests working directly with random functionals on Hilbert spaces. Finally, Anastasiou et al. (2021) provide a review of recent developments based on Stein's methods.

C Asymptotic level for wild bootstrap

The proof of asymptotic level using a wild bootstrap was proved by Chwialkowski et al. (2016, Proposition 3.2) relying on the result of Leucht (2012, Theorem 2.1) and of and Chwialkowski et al. (2014, Lemma 5). Their results for V-statistics can be extended to the use of U-statistics. They proved this result for the more general non-i.i.d. case while we need only the result for the simpler i.i.d. case. They proved that the difference between true quantiles and the wild bootstrap quantiles converges to zero in probability under the null hypothesis with the following dependence on N:

$$\sup_{t \in \mathbb{R}} \Big| \mathbb{P}(NB_N > t \mid \mathbb{X}_N) - \mathbb{P}(NK_N > t) \Big|$$

converges to 0 in probability, where K_N is the KSD estimator of Equation (1), and B_N is the wild bootstrap KSD of Equation (5). Lipschitz continuity of $h_{p,k}$ is required for the result to hold.

D Runtimes comparison of all KSD tests considered

The time complexity of KSDAGG is $\mathcal{O}\big(|\mathcal{K}|(B_1 + B_2)\, N^2\big)$ as provided in Algorithm 1. It grows linearly with the number of kernels $|\mathcal{K}|$, quadratically with the sample size N, and linearly with the number of bootstrap samples $B_1 + B_2$.

We report in Table 1 the runtimes of KSDAGG, KSD median, KSD split, and KSD split extra data, all using a wild bootstrap to estimate the quantiles in the MNIST Normalizing Flow experimental setting considered in Section 4.5.

Table 1: Runtimes (averaged over 10 runs and reported in seconds) for the Normalizing Flow MNIST experiment presented in Section 4.5 using a wild bootstrap.

SAMPLE SIZE	KSDAGG	KSD MEDIAN	KSD SPLIT	KSD SPLIT EXTRA DATA
100	0.037	0.005	0.022	0.023
200	0.084	0.010	0.064	0.070
300	0.162	0.020	0.132	0.145
400	0.276	0.034	0.230	0.253
500	0.421	0.051	0.359	0.395

Since we consider a collection of bandwidths Λ, the time complexity of KSDAGG is $\mathcal{O}(|\Lambda|(B_1 + B_2)N^2)$ and the one for median KSD is $\mathcal{O}(B_1 N^2)$ where $|\Lambda| = 21$ and $B_1 = B_2 = 500$. While KSDAGG takes roughly 10 times longer to run than KSD median, we could have expected a larger difference looking at the time complexities. This can be explained by the fact that there are two major time-consuming steps: (i) computing the kernel matrices, and (ii) computing the wild bootstrap samples. While (i) has complexity $\mathcal{O}(N^2)$ and (ii) complexity $\mathcal{O}(BN^2 + NB^2)$, the constant for step (i) is much larger than the one for step (ii) since (i) requires computing the matrix of pairwise distances while (ii) only consists in computing some matrix multiplications. Note that to compute the $|\Lambda|$ kernel matrices for KSDAGG, we need to compute the matrix of pairwise distances only once.

When splitting the data, the computationally expensive step is to select the bandwidth. All the $|\Lambda|$ kernel matrices need to be computed, as for KSDAGG, which is the expensive step (i). The KSD split test runs only slightly faster than the KSD split extra test; it runs faster than KSDAGG but their runtimes are of the same order of magnitude.

E Details on MNIST Normalizing Flow experiment

We provide details about the results in Figure 3b of the MNIST Normalizing Flow experiment presented in Section 4.5. In that case, the median bandwidth λ_{med} is on average 2437. The collection of bandwidths considered for KSDAGG is $\Lambda(-20, 0) := \left\{2^i \lambda_{\text{med}} : i = -20, \ldots, 0\right\}$. When KSDAGG rejects the null hypothesis, the smallest bandwidth $2^{-20}\lambda_{\text{med}} \approx 2^{-20} \cdot 2437 \approx 0.002$ (among others) rejects the single test with adjusted level. The bandwidth selected by KSD split extra data is the largest bandwidth $2^0 \lambda_{\text{med}} \approx 2437$. The bandwidth selection is performed by maximizing the criterion in Equation (9). Sutherland et al. (2017) showed that this is equivalent to maximizing asymptotic power in the case of the MMD; the same result holds straightforwardly for the KSD due to similar asymptotic properties. The criterion only maximizes asymptotic power and has no guarantee when using limited data. In this high-dimensional setting ($d = 784$) with sample sizes smaller than 500, the asymptotic regime is not attained, and the criterion used for bandwidth selection does not maximize the power in this non-asymptotic setting. So, even though split extra has access to some extra data, it does not have an accurate criterion to select the bandwidth and ends up selecting the largest bandwidth, which is not well-adapted to this problem. This explains the low power obtained by KSD split extra data.

F Details on our aggregation procedure

More powerful than Bonferroni correction. The Bonferroni correction for multiple testing corresponds to using the adjusted level $\alpha/|\mathcal{K}|$ for each of the $|\mathcal{K}|$ tests. With the correction used for KSDAGG with uniform weights, the adjusted level for each test is $u_\alpha/|\mathcal{K}|$ with u_α defined in Equation (6). It can be shown that $u_\alpha \geq \alpha$ (Albert et al., 2022, Lemma 4), which means that KSDAGG will always reject the null when the test with Bonferroni correction would reject it. When using uniform weights, the multiple testing correction we use is guaranteed to always lead to a test as least as powerful as the one using the Bonferroni correction. The Bonferroni correction uses a loose union bound argument in order to guarantee control of the probability of type I error, the method we use aims to tighten this loose upper bound. We present edge-case examples to illustrate the strengths of our multiple testing correction. First, assume that the $|\mathcal{K}|$ events are all disjoints, then the union

bound is tight and both methods yield the adjusted levels $\alpha/|\mathcal{K}|$. Second, assume that all events are the same (*e.g.* same kernels), then the Bonferroni correction still yields adjusted levels $\alpha/|\mathcal{K}|$, while our multiple testing strategy can detect that there is nothing to correct for and provide 'adjusted' levels α. We note that, given some fixed weights and level, there does not exist a single u_α associated to them, it will depend on the events.

Choice of weights. Without any prior knowledge (which is often the case in practice), we recommend using uniform weights since we do not expect any particular kernels/bandwidths to be better-suited than others. If the user has some prior knowledge of which kernels/bandwidths would be better for the task considered, then those should be given higher weights. We allow for weights whose sum is strictly smaller than 1 simply for the convenience of being able to add a new kernel/bandwidth with a new weight without changing the previous weights (as in Theorem 3.5 with weights $6/\pi^2\ell^2$ for $\ell \in \mathbb{N} \setminus \{0\}$). Multiplying all the weights by a constant simply results in dividing the correction u_α defined in Equation (6) by the same constant. This means that the product $u_\alpha w_\lambda$ remains the same, and hence the definition of the aggregated test is not affected by this change. For simplicity, in practice, we use weights whose sum is equal to 1.

Interpretability. When KSDAGG rejects the null hypothesis, we can check which specific kernels reject the adjusted tests: this provides the user with information about the kernels which are well-adapted to the problem considered. The 'best' selection of kernels is naturally returned as a side-effect of the test (without requiring data splitting). This contributes to the interpretability of KSDAGG, for instance in the case where different kernels prioritise different features.

Extension to uncountable collections of kernels. Our KSDAGG test aggregates over a finite collection of kernels. While even for large collections of kernels KSDAGG retains high power (due our multiple testing correction), it would be beneficial to be able to consider an uncountable/continuous collection of kernels (parametrised by the kernel bandwidth on the positive real line). Optimizing the kernel parameters continuously can be done by selecting the parameters which maximise a proxy for test power, but to the best of our knowledge, this has currently never been done without data splitting (which usually result in a loss of power). The extension to the case of uncountable collections remains a very challenging problem, and it would be of great interest to derive such a method in the future.

Extension to KSD tests with adaptive features. Our proposed test, KSDAGG, proposes a solution to the problem of KSD adaptivity in the goodness-of-fit framework without requiring data splitting. A potential future direction of interest could be to tackle the adaptivity problem of the KSD-based linear-time goodness-of-fit test proposed by Jitkrittum et al. (2017). In this setting, the data is split to select feature locations (and the kernel bandwidth), the KSD test is then run using those adaptive features. A challenging problem would be to obtain those adaptive features using an aggregation procedure which avoids splitting the data.

G Adversarially constructed settings in which KSDAGG might fail

Median bandwidth test. In settings (possibly adversarially constructed) where the median bandwidth is the 'best' bandwidth, the KSD test with median bandwidth could in theory be more competitive than KSDAGG since by considering a large collection of bandwidths we are not only considering the 'best' median bandwidth but also 'worse' bandwidths. However, in practice, we cannot know in advance which bandwidth would perform well, and KSDAGG retains power even for large collections of bandwidths (21 bandwidths considered in MNIST Normalizing Flow experiment in Section 4.5).

Data splitting test. Another setting (also possibly adversarially constructed) could be one in which the 'best' bandwidth lies in between two bandwidths of our collection and those two are 'worse' bandwidths. In that case, a test which uses data splitting to select an 'optimal' bandwidth would be able to select it, however one must bear in mind the loss of power due to data splitting. In such a setting, the power of KSDAGG could then be improved by considering a finer collection of bandwidths. However, in our experiments, the aggregation approach outperformed competing approaches.

H Proofs

H.1 Proof of Theorem 3.1

Note that

$$\begin{aligned}\mathbb{P}_q\big(\Delta_\alpha^k(\mathbb{X}_N) \;=\; 0\big) \;&=\; \mathbb{P}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) \;\leq\; \widehat{q}^k_{1-\alpha}\Big)\\ &=\; \mathbb{P}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) - \mathrm{KSD}^2_{p,k}(q) \;\leq\; \widehat{q}^k_{1-\alpha} - \mathrm{KSD}^2_{p,k}(q)\Big)\end{aligned}$$

where

$$\mathrm{KSD}^2_{p,k}(q) \;=\; \mathbb{E}_q\Big[\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big]$$

since $\widehat{\mathrm{KSD}}^2_{p,k}$ is an unbiased estimator. By Chebyshev's inequality (Chebyshev, 1899), we know that

$$\begin{aligned}\beta \;\geq\; \mathbb{P}_q\left(\left|\mathrm{KSD}^2_{p,k}(q) - \widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\right| \;\geq\; \sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}}\right)\\ \geq\; \mathbb{P}_q\left(\mathrm{KSD}^2_{p,k}(q) - \widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) \;\geq\; \sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}}\right)\\ =\; \mathbb{P}_q\left(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) - \mathrm{KSD}^2_{p,k}(q) \;\leq\; -\sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}}\right).\end{aligned}$$

We deduce that $\mathbb{P}_q\big(\Delta_\alpha^k(\mathbb{X}_N) = 0\big) \leq \beta$ if

$$\begin{aligned}\widehat{q}^k_{1-\alpha} - \mathrm{KSD}^2_{p,k}(q) \;&\leq\; -\sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}}\\ \mathrm{KSD}^2_{p,k}(q) \;&\geq\; \widehat{q}^k_{1-\alpha} + \sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}}.\end{aligned} \tag{10}$$

The condition in Equation (10), which controls the probability of type I error of Δ_α^k, needs to hold with high probability since $\widehat{q}^k_{1-\alpha}$ depends on the randomness of the samples $\mathbb{X}_N$ drawn from q and on the randomness of the Rademacher random variables in Equation (5).

We want to derive a condition in terms of $\|p-q\|_2^2$ rather than in terms of $\mathrm{KSD}^2_{p,k}(q)$ as in Equation (10). For this, using Stein's identity, we obtain

$$\begin{aligned}\big\langle \psi, T_{h_{p,k}}\psi\big\rangle_2 &= \int_{\mathbb{R}^d} \psi(y)\big(T_{h_{p,k}}\psi\big)(y)\,\mathrm{d}y\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{p,k}(x,y)\psi(x)\psi(y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{p,k}(x,y)(p(x)-q(x))(p(y)-q(y))\;\mathrm{d}x\,\mathrm{d}y\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{p,k}(x,y)q(x)q(y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \mathbb{E}_{q,q}[h_{p,k}(X,Y)]\\ &= \mathrm{KSD}^2_{p,k}(q)\end{aligned}$$

which gives

$$\mathrm{KSD}^2_{p,k}(q) = \langle \psi, T_{h_{p,k}}\psi\rangle = \frac{1}{2}\Big(\|\psi\|_2^2 + \big\|T_{h_{p,k}}\psi\big\|_2^2 - \big\|\psi - T_{h_{p,k}}\psi\big\|_2^2\Big).$$

To guarantee $\mathbb{P}_q\big(\Delta_\alpha^k(\mathbb{X}_N) = 0\big) \leq \beta$, an equivalent condition to the one presented in Equation (10) is then

$$\|\psi\|_2^2 \;\geq\; \big\|\psi - T_{h_{p,k}}\psi\big\|_2^2 - \big\|T_{h_{p,k}}\psi\big\|_2^2 + 2\widehat{q}_{1-\alpha}^k + 2\sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}_{p,k}^2(\mathbb{X}_N)\Big)}{\beta}} \tag{11}$$

which needs to hold with high probability over the randomness in $\widehat{q}_{1-\alpha}^k$. We now upper bound the quantile and variance terms in Equation (11) to obtain a new condition ensuring control of the type II error.

To bound the quantile term $\widehat{q}_{1-\alpha}^k$ using B_1 wild bootstrapped statistics, using the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990) as done by Schrab et al. (2021, Appendix E.4), it is sufficient to bound the quantile $\widehat{q}_{1-\alpha}^{k,\infty}$ using wild bootstrapped statistics without finite approximation provided that $B_1 \geq \frac{3}{\alpha^2}\big(\log\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$. Using the concentration bound for i.i.d. Rademacher chaos of de la Peña and Giné (1999, Corollary 3.2.6) as presented by Kim et al. (2022, Equation 39), there exists some constant $C > 0$ such that

$$\mathbb{P}_\epsilon\left(\left|\frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N}\epsilon_i\epsilon_j h_{p,k}(X_i,X_j)\right| \geq t \;\middle|\; \mathbb{X}_N\right) \leq 2\exp\left(-\frac{CtN(N-1)}{\sqrt{\sum_{1\leq i\neq j\leq N} h_{p,k}(X_i,X_j)^2}}\right)$$

which gives

$$\widehat{q}_{1-\alpha}^{k,\infty} \;\leq\; C\frac{\log(2/\alpha)}{N(N-1)}\sqrt{\sum_{1\leq i\neq j\leq N} h_{p,k}(X_i,X_j)^2}$$

for some different constant $C > 0$. Using Markov's inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$\sum_{1\leq i\neq j\leq N} h_{p,k}(X_i,X_j)^2 \leq \frac{1}{\delta}N(N-1)\mathbb{E}_{q,q}\big[h_{p,k}(X,Y)^2\big] = \frac{C_k N(N-1)}{\delta}$$

which gives

$$\widehat{q}_{1-\alpha}^{k,\infty} \leq C\frac{\log(2/\alpha)}{N(N-1)}\sqrt{\frac{C_k N(N-1)}{\delta}} = \frac{C}{\sqrt{\delta}}\frac{\log(2/\alpha)}{\sqrt{N(N-1)}}\sqrt{C_k} \leq C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{C_k}}{N}$$

since $\alpha \in (0, e^{-1})$, where in the last inequality $C > 0$ is a different constant depending on δ. Using Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956) as explained above, we deduce that

$$2\widehat{q}_{1-\alpha}^k \;\leq\; C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{C_k}}{N} \tag{12}$$

with arbitrarily high probability, for some constant $C > 0$. The same quantile bound also holds for the parametric bootstrap provided that $n/\sqrt{C_k} \geq \ln(1/\alpha)$ as derived by Albert et al. (2022, Appendix C.4.1). Their reasoning holds in our setting by replacing $1/\sqrt{\lambda_1\cdots\lambda_p\mu_1\cdots\mu_q}$ by $\sqrt{C_k}$, this is justified as for a kernel k_λ with bandwidths $\lambda_1,\ldots,\lambda_d$ we have $\mathbb{E}_{q,q}[k_\lambda(X,Y)^2] \leq C/\lambda_1\cdots\lambda_d$.

We now bound the variance term in Equation (10). Using the result of Albert et al. (2022, Equation 6), there exists $C > 0$ such that

$$\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}_{p,k}^2(\mathbb{X}_N)\Big) \leq C\left(\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N^2}\right)$$

where

$$\sigma_2^2 \coloneqq \mathbb{E}_{q,q}\big[h_{p,k}(X,Y)^2\big] = C_k$$

and

$$\begin{aligned}
\sigma_1^2 &\coloneqq \mathbb{E}_{Y\sim q}\left[\Big(\mathbb{E}_{X\sim q}\left[h_{p,k}(X,Y)\right]\Big)^2\right] \\
&= \mathbb{E}_{Y\sim q}\left[\Big(\big(T_{h_{p,k}}\psi\big)(Y)\Big)^2\right] \\
&= \int_{\mathbb{R}^d}\Big(\big(T_{h_{p,k}}\psi\big)(y)\Big)^2 q(y)\,\mathrm{d}y \\
&\leq \|q\|_\infty \int_{\mathbb{R}^d}\Big(\big(T_{h_{p,k}}\psi\big)(y)\Big)^2 \mathrm{d}y \\
&\leq M\big\|T_{h_{p,k}}\psi\big\|_2^2
\end{aligned}$$

since

$$\begin{aligned}
\big(T_{h_{p,k}}\psi\big)(y) &\coloneqq \int_{\mathbb{R}^d} h_{p,k}(x,y)\psi(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} h_{p,k}(x,y)p(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} h_{p,k}(x,y)q(x)\,\mathrm{d}x \\
&= 0 - \int_{\mathbb{R}^d} h_{p,k}(x,y)q(x)\,\mathrm{d}x \\
&= -\mathbb{E}_{X\sim q}\left[h_{p,k}(X,y)\right]
\end{aligned}$$

by Stein's identity. We deduce that

$$\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big) \leq C\left(\frac{\big\|T_{h_{p,k}}\psi\big\|_2^2}{N} + \frac{C_k}{N^2}\right)$$

for some different constant $C > 0$ depending on M.

Using the classical inequalities $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and $2\sqrt{xy} \leq x + y$ for $x, y > 0$, which are also considered in the works of Fromont et al. (2013), Albert et al. (2022) and Schrab et al. (2021), we obtain

$$\begin{aligned}
2\sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}} &\leq 2\sqrt{C\frac{\big\|T_{h_{p,k}}\psi\big\|_2^2}{\beta N} + C\frac{C_k}{\beta N^2}} \\
&\leq 2\sqrt{\big\|T_{h_{p,k}}\psi\big\|_2^2\frac{C}{\beta N}} + 2\sqrt{C\frac{C_k}{\beta N^2}} \\
&\leq \big\|T_{h_{p,k}}\psi\big\|_2^2 + \frac{C}{\beta N} + 2\sqrt{C}\frac{\sqrt{C_k}}{\sqrt{\beta}N} \\
&\leq \big\|T_{h_{p,k}}\psi\big\|_2^2 + \Big(C + 2\sqrt{C}\Big)\frac{\sqrt{C_k}}{\beta N} \\
&\leq \big\|T_{h_{p,k}}\psi\big\|_2^2 + C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{C_k}}{\beta N} \qquad (13)
\end{aligned}$$

since $\alpha \in (0, e^{-1})$, $\beta \in (0,1)$ and $C_k \geq 1$ by assumption, where the constant $C > 0$ is different on the last line. By considering two times the largest of the two constants in the quantile and variance bounds of Equations (12) and (13), we obtain

$$2\widehat{q}^k_{1-\alpha} + 2\sqrt{\frac{\mathrm{var}_q\Big(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)\Big)}{\beta}} \;\leq\; \big\|T_{h_{p,k}}\psi\big\|_2^2 + C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{C_k}}{\beta N}.$$

By combining this bound with the condition in Equation (11), we get that $\mathbb{P}_q\big(\Delta^k_\alpha(\mathbb{X}_N) = 0\big) \leq \beta$ if

$$\begin{aligned}
\|\psi\|_2^2 \;&\geq\; \big\|\psi - T_{h_{p,k}}\psi\big\|_2^2 - \big\|T_{h_{p,k}}\psi\big\|_2^2 + \big\|T_{h_{p,k}}\psi\big\|_2^2 + C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{C_k}}{\beta N}, \\
\|\psi\|_2^2 \;&\geq\; \big\|\psi - T_{h_{p,k}}\psi\big\|_2^2 + C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{C_k}}{\beta N},
\end{aligned}$$

which concludes the proof.

H.2 Proof of Proposition 3.2

Recall that the correction term in Equation (6) is defined as

$$u_\alpha \coloneqq \sup\left\{u \in \left(0, \min_{k\in\mathcal{K}} w_k^{-1}\right) : \frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{k\in\mathcal{K}}\left(\widetilde{K}_k^b - \widehat{q}^k_{1-uw_k}\right) > 0\right) \leq \alpha\right\}$$

where $\widehat{q}^k_{1-uw_k} = \bar{K}_k^{\bullet\lceil B_1(1-uw_k)\rceil}$ as defined in Equation (3). Hence, we have

$$\frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}\left(\max_{k\in\mathcal{K}}\left(\widetilde{K}_k^b - \widehat{q}^k_{1-u_\alpha w_k}\right) > 0\right) \;\leq\; \alpha. \tag{14}$$

Recall that our estimator is $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)$ where $\mathbb{X}_N \coloneqq (X_i)_{i=1}^N$ are drawn from q. Recall also from Equation (4), that for the parametric bootstrap the element $\widetilde{K}_k^b$ is computed by drawing new samples $(X_i')_{i=1}^{N'}$ from the model p and computing $\widehat{\mathrm{KSD}}^2_{p,k}\left((X_i')_{i=1}^{N'}\right)$. Hence, at every sample size, under the null hypothesis $\mathcal{H}_0 : p = q$ the quantities $\widetilde{K}_k^b$ and $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)$ are identically distributed. For the wild bootstrap, as defined in Equation (5), Chwialkowski et al. (2014, 2016) show that, under $\mathcal{H}_0$, $\widetilde{K}_k^b$ and $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N)$ have the same asymptotic distribution.

Therefore, by taking the expectation in Equation (14), we obtain

$$\begin{aligned}
\alpha \;&\geq\; \mathbb{P}_p\left(\max_{k\in\mathcal{K}}\left(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) - \widehat{q}^k_{1-u_\alpha w_k}\right) > 0\right)\\
&=\; \mathbb{P}_p\left(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{X}_N) > \widehat{q}^k_{1-u_\alpha w_k} \;\text{ for some }\; k\in\mathcal{K}\right)\\
&=\; \mathbb{P}_p\left(\Delta^k_{u_\alpha w_k}(\mathbb{X}_N) = 1 \;\text{ for some }\; k\in\mathcal{K}\right)\\
&=\; \mathbb{P}_p\left(\Delta^{\mathcal{K}}_\alpha(\mathbb{X}_N) = 1\right)
\end{aligned}$$

which holds non-asymptotically using the parametric bootstrap, and asymptotically using the wild bootstrap. Note that this is different from the two-sample case where using a wild bootstrap yields well-calibrated non-asymptotic levels (Schrab et al., 2021, Proposition 8).

The same reasoning holds by replacing u_α by any value $u \in (0, u_\alpha)$. In particular, it holds for the lower bound $\widehat{u}_\alpha$ of the interval obtained by performing the bisection method to approximate the supremum in the definition of u_α. This lower bound $\widehat{u}_\alpha$ is the one we use in practice, as shown in Algorithm 1 with the step '$u_\alpha = u_{\text{min}}$'. We have proved that the test with correction term u_α approximated with a bisection method also achieves the desired level α.

H.3 Proof of Theorem 3.3

As explained in Appendix H.2 and utilised in Algorithm 1, we use the lower bound $\widehat{u}_\alpha$ of the interval obtained by performing the bisection method to approximate the supremum in the definition of u_α in Equation (6). The assumptions $B_2 \geq \frac{8}{\alpha^2}\log\left(\frac{2}{\beta}\right)$ and $B_3 \geq \log_2\left(\frac{4}{\alpha}\min_{k\in\mathcal{K}} w_k^{-1}\right)$ ensure that $\widehat{u}_\alpha \geq \alpha/2$ as shown by Schrab et al. (2021, Appendix E.9).

The probability of type II error of $\Delta^{\mathcal{K}}_\alpha$ is

$$\begin{aligned}
\mathbb{P}_q\left(\Delta^{\mathcal{K}}_\alpha(\mathbb{X}_N) = 0\right) \;&=\; \mathbb{P}_q\left(\Delta^k_{\widehat{u}_\alpha w_k}(\mathbb{X}_N) = 0 \;\text{ for all }\; k\in\mathcal{K}\right)\\
&\leq\; \mathbb{P}_q\left(\Delta^k_{\widehat{u}_\alpha w_k}(\mathbb{X}_N) = 0 \;\text{ for some }\; k\in\mathcal{K}\right)\\
&\leq\; \mathbb{P}_q\left(\Delta^k_{\alpha w_k/2}(\mathbb{X}_N) = 0 \;\text{ for some }\; k\in\mathcal{K}\right).
\end{aligned}$$

In order to guarantee $\mathbb{P}_q\left(\Delta^{\mathcal{K}}_\alpha(\mathbb{X}_N) = 0\right) \leq \beta$ it is sufficient to guarantee $\mathbb{P}_q\left(\Delta^{k^*}_{\alpha w_{k^*}/2}(\mathbb{X}_N) = 0\right) \leq \beta$ for some $k^* \in \mathcal{K}$ to be specified shortly in Equation (15). By assumption, we have

$$B_1 \;\geq\; \left(\max_{k\in\mathcal{K}} w_k^{-2}\right)\frac{12}{\alpha^2}\left(\log\left(\frac{8}{\beta}\right) + \alpha(1-\alpha)\right).$$

In order to apply Theorem 3.1 to $\Delta^k_{\alpha_k}$ with $\alpha_k := \alpha w_k/2$ for $k \in \mathcal{K}$, we need to ensure that the condition on B_1 of Theorem 3.1 is satisfied for all $k \in \mathcal{K}$, that is

$$B_1 \;\geq\; \frac{3}{\alpha_k^2}\left(\log\left(\frac{8}{\beta}\right) + \alpha_k(1-\alpha_k)\right)$$

for $k \in \mathcal{K}$. Since $0 < \alpha_k < \alpha < e^{-1}$, we have $\alpha(1-\alpha) \geq \alpha_k(1-\alpha_k)$ for $k \in \mathcal{K}$. We also have

$$\left(\max_{k\in\mathcal{K}} w_k^{-2}\right)\frac{12}{\alpha^2} \;\geq\; 3\left(\frac{2}{\alpha w_k}\right)^2 = \frac{3}{\alpha_k^2}$$

for $k \in \mathcal{K}$. We deduce that

$$\begin{aligned} B_1 \;&\geq\; \left(\max_{k\in\mathcal{K}} w_k^{-2}\right)\frac{12}{\alpha^2}\left(\log\left(\frac{8}{\beta}\right) + \alpha(1-\alpha)\right) \\ &\geq\; \frac{3}{\alpha_k^2}\left(\log\left(\frac{8}{\beta}\right) + \alpha_k(1-\alpha_k)\right) \end{aligned}$$

for $k \in \mathcal{K}$, and so, applying Theorem 3.1 to $\Delta^k_{\alpha_k}$ for $k \in \mathcal{K}$ is justified. We obtain that $\mathbb{P}_q\left(\Delta^k_{\alpha w_k/2}(\mathbb{X}_N) = 0\right) \leq \beta$ if

$$\|\psi\|_2^2 \;\geq\; \left\|\psi - T_{h_{p,k}}\psi\right\|_2^2 + C\log\left(\frac{2}{\alpha w_k}\right)\frac{\sqrt{C_k}}{\beta N},$$

or, with a different constant $C > 0$, if

$$\|\psi\|_2^2 \;\geq\; \left\|\psi - T_{h_{p,k}}\psi\right\|_2^2 + C\log\left(\frac{1}{\alpha w_k}\right)\frac{\sqrt{C_k}}{\beta N}$$

since $\log\left(\frac{2}{\alpha w_k}\right) \leq (\log(2)+1)\log\left(\frac{1}{\alpha w_k}\right)$ as $\alpha \in (0, e^{-1})$ and $w_k \in (0,1)$. Now, let

$$k^* := \operatorname*{argmin}_{k\in\mathcal{K}}\left(\left\|\psi - T_{h_{p,k}}\psi\right\|_2^2 + C\log\left(\frac{1}{\alpha w_k}\right)\frac{\sqrt{C_k}}{\beta N}\right). \tag{15}$$

Finally, we have $\mathbb{P}_q\left(\Delta^{\mathcal{K}}_{\alpha}(\mathbb{X}_N) = 0\right) \leq \beta$ if $\mathbb{P}_q\left(\Delta^{k^*}_{\alpha w_{k^*}/2}(\mathbb{X}_N) = 0\right) \leq \beta$, that is, if

$$\begin{aligned} \|\psi\|_2^2 \;&\geq\; \left\|\psi - T_{h_{p,k^*}}\psi\right\|_2^2 + C\log\left(\frac{1}{\alpha w_{k^*}}\right)\frac{\sqrt{C_{k^*}}}{\beta N} \\ &=\; \min_{k\in\mathcal{K}}\left(\left\|\psi - T_{h_{p,k}}\psi\right\|_2^2 + C\log\left(\frac{1}{\alpha w_k}\right)\frac{\sqrt{C_k}}{\beta N}\right), \end{aligned}$$

as desired.

H.4 Proof of Theorem 3.5

Recall that we suppose that the following assumptions hold.

- The model density p is strictly positive on its connected and compact support $S \subseteq \mathbb{R}^d$.
- The score function $\nabla \log p(x)$ is continuous and bounded on S.
- The support of the density q is a connected and compact subset of S.
- The kernel used is a scaled Gaussian kernel $k_\lambda(x,y) := \lambda^{2-d}\exp\left(-\|x-y\|_2^2 / \lambda^2\right)$.

We introduce some notation. For some $c \in \mathbb{R}$, we write $\mathbf{c} := (c, \ldots, c) \in \mathbb{R}^d$. We write $(a_1, \ldots, a_d) \leq (b_1, \ldots, b_d)$ when $a_i \leq b_i$ for $i = 1, \ldots, d$. We use C, C' to denote some generic constants which may change on different lines.

Note that, by properties of compactness on $\mathbb{R}^d$, there exists some $a > 0$ such that $S \subseteq \left[-\frac{a}{2}, \frac{a}{2}\right]^d$. Since the score function $\nabla \log p(x)$ is continuous and bounded on S which is compact, there exists some $c_1 > 0$ such that

$$|(\nabla \log p(x))_i| \leq c_1 \quad \text{for } i = 1, \ldots, d \text{ and for all } x \in S.$$

We work with a scaled Gaussian kernel with bandwidth $\lambda \leq 1$

$$k_\lambda(x,y) \coloneqq \lambda^{2-d} \exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right)$$

which satisfies

$$\begin{aligned}
\nabla_x k_\lambda(x,y) &= 2\lambda^{-d}(y-x)\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right) \leq 2\lambda^{-d}\mathbf{a}\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right), \\
\nabla_y k_\lambda(x,y) &= 2\lambda^{-d}(x-y)\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right) \leq 2\lambda^{-d}\mathbf{a}\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right), \\
\sum_{i=1}^d \frac{\partial}{\partial x_i \partial y_i} k_\lambda(x,y) &= \left(2d\lambda^{-d} - 4\lambda^{-2d}\|x-y\|_2^2\right)\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right) \\
&\leq 2d\lambda^{-d}\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right).
\end{aligned}$$

The Stein kernel associated to k_λ satisfies

$$\begin{aligned}
h_{p,k_\lambda}(x,y) &\coloneqq \left(\nabla \log p(x)^\top \nabla \log p(y)\right) k_\lambda(x,y) + \nabla \log p(y)^\top \nabla_x k_\lambda(x,y) \\
&\quad + \nabla \log p(x)^\top \nabla_y k_\lambda(x,y) + \sum_{i=1}^d \frac{\partial}{\partial x_i \partial y_i} k_\lambda(x,y) \\
&\leq 2\left(dc_1^2 + 2dc_1 a\lambda^{-d} + d\lambda^{-d}\right)\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right) \\
&\leq C_0 \pi^{-d/2}\lambda^{-d}\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right)
\end{aligned}$$

where the constant $C_0 \coloneqq 2\pi^{d/2}(dc_1^2 + 2dc_1 a + d)$. The Stein kernel can be upper bounded by a scaled Gaussian kernel on S. Writing

$$\bar{k}_\lambda(x,y) \coloneqq \pi^{-d/2}\lambda^{-d}\exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right),$$

which is of the form considered by Schrab et al. (2021, Section 3.1) with equal bandwidths for each dimension, we have shown that

$$h_{p,k_\lambda}(x,y) \leq C_0 \bar{k}_\lambda(x,y) \qquad \text{for all } x,y \in S. \tag{16}$$

Writing $\psi \coloneqq p - q$, recall from Theorem 3.1 that a sufficient condition for control of the probability of type II error by β is

$$\|\psi\|_2^2 \geq \left\|\psi - T_{h_{p,k_\lambda}}\psi\right\|_2^2 + C\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{\mathbb{E}_{q,q}[h_{p,k_\lambda}(X,Y)^2]}}{\beta N}.$$

Using the upper bound

$$\mathbb{E}_{q,q}[h_{p,k_\lambda}(X,Y)^2] \leq C_0^2 \mathbb{E}_{q,q}\left[\bar{k}_\lambda(X,Y)^2\right] \leq C_0^2 \frac{M}{\lambda^d}$$

where the last inequality holds as in Schrab et al. (2021, Equation (22)), we obtain the sufficient condition

$$\|\psi\|_2^2 \geq \left\|\psi - T_{h_{p,k_\lambda}}\psi\right\|_2^2 + C\frac{\log(1/\alpha)}{\beta N \lambda^{d/2}} \tag{17}$$

where we have absorbed the term $C_0\sqrt{M}$ in the constant C.

We would like to upper bound the term $\left\|\psi - T_{h_{p,k_\lambda}}\psi\right\|_2^2$ by $\left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2$ but this is not possible using Equation (16). Instead, we can use the triangle inequality to get

$$\begin{aligned}\left\|\psi - T_{h_{p,k_\lambda}}\psi\right\|_2^2 &\leq \left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2 + \left\|T_{\bar{k}_\lambda}\psi - T_{h_{p,k_\lambda}}\psi\right\|_2^2 \\ &\leq \left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2 + 2\left(\left\|T_{\bar{k}_\lambda}\psi\right\|_2^2 + \left\|T_{h_{p,k_\lambda}}\psi\right\|_2^2\right) \\ &\leq \left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2 + 2(C_0^2+1)\left\|T_{\bar{k}_\lambda}\psi\right\|_2^2\end{aligned}$$

since

$$\big(T_{h_{p,k_\lambda}}\psi\big)(y) = \int_S h_{p,k_\lambda}(x,y)\psi(x)\,\mathrm{d}x \leq C_0 \int_S \bar{k}_\lambda(x,y)\psi(x)\,\mathrm{d}x = C_0\big(T_{\bar{k}_\lambda}\psi\big)(y)$$

for all $y \in S$.

Recall that we assume that $\psi \in \mathcal{S}_d^{s,t}(R,L)$ for some L to be determined, in particular $\psi \in \mathcal{S}_d^s(R)$.

For the term $\left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2$, we use the fact that $\psi \in \mathcal{S}_d^s(R)$. The term $\left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2$ can then be upper bounded exactly as done by Schrab et al. (2021, Appendix E.6) with the difference that we choose $\widetilde{t} > 0$ such that $S_1 < 0.5$ rather than $S_1 < 1$ (S_1 is defined in Schrab et al. (2021, Appendix E.6)). Following their reasoning, assuming that $\psi \in \mathcal{S}_d^s(R)$, we obtain that there exists some $S_1 \in (0, 0.5)$ and some constant $C > 0$ (depending on d, s, R) such that

$$\left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2 \leq S_1^2\|\psi\|_2^2 + C\lambda^{2s} \leq \frac{1}{4}\|\psi\|_2^2 + C\lambda^{2s}.$$

To upper bound the term $\left\|T_{\bar{k}_\lambda}\psi\right\|_2^2$, we modify the reasoning of Schrab et al. (2021, Appendix E.6) used for the term $\left\|\psi - T_{\bar{k}_\lambda}\psi\right\|_2^2$. First, note that since $\bar{k}_\lambda$ is translation-invariant, $T_{\bar{k}_\lambda}$ is a convolution as

$$\big(T_{\bar{k}_\lambda}\psi\big)(y) = \int_S \bar{k}_\lambda(x,y)\psi(x)\,\mathrm{d}x = \int_S \varphi_\lambda(x-y)\psi(x)\,\mathrm{d}x = \big(\psi * \varphi_\lambda\big)(y)$$

where

$$\varphi_\lambda(u) \coloneqq \prod_{i=1}^d \lambda^{-1} K\left(\frac{u_i}{\lambda}\right), \qquad \text{and} \qquad K(u) \coloneqq \pi^{-1/2}\exp(-u^2).$$

By properties of Fourier transforms, we have $\widehat{\varphi_\lambda}(\xi) = \prod_{i=1}^d \widehat{K}(\lambda\xi_i)$ for $\xi \in \mathbb{R}^d$. Note that

$$\left|\prod_{i=1}^d \widehat{K}(\xi_i)\right| \leq \prod_{i=1}^d \int_{\mathbb{R}} \left|K(x)e^{-ix\xi_i}\right|\mathrm{d}x = \prod_{i=1}^d \int_{\mathbb{R}} |K(x)|\,\mathrm{d}x = 1.$$

We now let $L \coloneqq 8(C_0^2+1)$. By assumption, we have $\psi \in \mathcal{S}_d^{s,t}(R,L)$, so

$$\int_{\|\xi\|_2 \leq t} \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi \leq \frac{1}{L}\int_{\mathbb{R}^d} \left|\widehat{\psi}(\xi)\right|^2 \mathrm{d}\xi = \frac{\left\|\widehat{\psi}\right\|_2^2}{L}.$$

For $s > 0$, define

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

Then, using Plancherel's Theorem as in Schrab et al. (2021, Appendix E.6), we obtain

$$
\begin{aligned}
&(2\pi)^d \left\| T_{\bar{k}_\lambda} \psi \right\|_2^2 \\
&= (2\pi)^d \|\psi * \varphi_\lambda\|_2^2 \\
&= \left\| \widehat{\varphi_\lambda} \widehat{\psi} \right\|_2^2 \\
&= \int_{\mathbb{R}^d} |\widehat{\varphi_\lambda}(\xi)|^2 \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi \\
&= \int_{\mathbb{R}^d} \left| \prod_{i=1}^d \widehat{K}(\lambda \xi_i) \right|^2 \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi \\
&= \int_{\|\xi\|_2 \leq t} \left| \prod_{i=1}^d \widehat{K}(\lambda \xi_i) \right|^2 \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi + \int_{\|\xi\|_2 > t} \left| \prod_{i=1}^d \widehat{K}(\lambda \xi_i) \right|^2 \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi \\
&\leq \int_{\|\xi\|_2 \leq t} \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi + T_s^2 \int_{\|\xi\|_2 > t} \|\lambda \xi\|_2^{2s} \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi \\
&\leq \frac{1}{L} \left\| \widehat{\psi} \right\|_2^2 + \lambda^{2s} T_s^2 \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \left| \widehat{\psi}(\xi) \right|^2 \mathrm{d}\xi \\
&\leq \frac{(2\pi)^d}{L} \|\psi\|_2^2 + \lambda^{2s} T_s^2 (2\pi)^d R^2.
\end{aligned}
$$

Combining those results, the upper bound on $\left\| \psi - T_{h_{p,k_\lambda}} \psi \right\|_2^2$ becomes

$$
\begin{aligned}
\left\| \psi - T_{h_{p,k_\lambda}} \psi \right\|_2^2 &\leq \left\| \psi - T_{\bar{k}_\lambda} \psi \right\|_2^2 + (C_0 + 1)^2 \left\| T_{\bar{k}_\lambda} \psi \right\|_2^2 \\
&\leq \left(\frac{1}{4} \|\psi\|_2^2 + C \lambda^{2s} \right) + \left(\frac{2(C_0^2 + 1)}{L} \|\psi\|_2^2 + C' \lambda^{2s} \right) \\
&= \left(\frac{1}{4} \|\psi\|_2^2 + C \lambda^{2s} \right) + \left(\frac{1}{4} \|\psi\|_2^2 + C' \lambda^{2s} \right) \\
&\leq \frac{1}{2} \|\psi\|_2^2 + C \lambda^{2s}.
\end{aligned}
$$

Equation (17) then becomes

$$
\|\psi\|_2^2 \geq \frac{1}{2} \|\psi\|_2^2 + C \lambda^{2s} + C' \frac{\log(1/\alpha)}{\beta N \lambda^{d/2}}
$$

which gives

$$
\|\psi\|_2^2 \geq C \left(\lambda^{2s} + \frac{\log(1/\alpha)}{\beta N \lambda^{d/2}} \right). \tag{18}
$$

Having proved the power guaranteeing condition in Equation (18), we can then derive rates for the KSD and KSDAGG tests as similarly done by Schrab et al. (2021). Set $\lambda := N^{-2/(4s+d)}$ to get the condition

$$
\|\psi\|_2^2 \geq C \left(N^{-4s/(4s+d)} + N^{-1} N^{d/(4s+d)} \right) = C N^{-4s/(4s+d)}
$$

which is the minimax rate over (unrestricted) Sobolev balls $\mathcal{S}_d^s(R)$ (Li and Yuan, 2019; Balasubramanian et al., 2021; Albert et al., 2022; Schrab et al., 2021).

In practice, the smoothness parameter of the Sobolev ball is not known, so we cannot set $\lambda := N^{-2/(4s+d)}$ (i.e. it cannot be implemented). Instead, we can use our aggregated test KSDAGG. Adapting the proof of Theorem 3.3 with the single condition Equation (18), we obtain the aggregated condition

$$
\|\psi\|_2^2 \geq C \min_{\lambda \in \Lambda} \left(\lambda^{2s} + \log\left(\frac{1}{\alpha w_\lambda} \right) \frac{1}{\beta N \lambda^{d/2}} \right). \tag{19}
$$

Similarly to Schrab et al. (2021, Corollary 10), consider

$$\Lambda \coloneqq \Big\{2^{-\ell} : \ell \in \Big\{1, \dots, \Big\lceil \frac{2}{d} \log_2\Big(\frac{N}{\ln(\ln(N))}\Big)\Big\rceil\Big\}\Big\}, \qquad w_\lambda \coloneqq \frac{6}{\pi^2\,\ell^2}.$$

Let

$$\ell_* \coloneqq \Big\lceil \frac{2}{4s+d} \log_2\Big(\frac{N}{\ln(\ln(N))}\Big)\Big\rceil \leq \Big\lceil \frac{2}{d} \log_2\Big(\frac{N}{\ln(\ln(N))}\Big)\Big\rceil.$$

The bandwidth $\lambda_* \coloneqq 2^{-\ell_*} \in \Lambda$ satisfies

$$\ln\Big(\frac{1}{w_{\lambda_*}}\Big) \leq C \ln(\ell_*) \leq C \ln(\ln(N))$$

as $w_{\lambda_*} \coloneqq 6\pi^{-2}(\ell_*)^{-2}$, and

$$\frac{1}{2}\Big(\frac{N}{\ln(\ln(N))}\Big)^{-2/(4s+d)} \leq \lambda_* \leq \Big(\frac{N}{\ln(\ln(N))}\Big)^{-2/(4s+d)}.$$

We get

$$\lambda_*^{2s} \leq \Big(\frac{N}{\ln(\ln(N))}\Big)^{-4s/(4s+d)}$$

and

$$\log\Big(\frac{1}{\alpha w_{\lambda_*}}\Big)\frac{1}{\beta N \lambda_*^{d/2}} \leq C\Big(\frac{N}{\ln(\ln(N))}\Big)^{-1}\Big(\frac{N}{\ln(\ln(N))}\Big)^{d/(4s+d)} \leq C\Big(\frac{N}{\ln(\ln(N))}\Big)^{-4s/(4s+d)}.$$

The power condition Equation (19) becomes

$$\|\psi\|_2^2 \geq C\Big(\frac{N}{\ln(\ln(N))}\Big)^{-4s/(4s+d)},$$

which concludes the proof.