KSD Aggregated Goodness-of-fit Test

Anonymous Author(s)
Affiliation
Address
email

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. 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 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 we are interested in deciding 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 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

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

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. Finally, Key et al. (2021) addressed the complementary task of Stein test design for a family of models, rather than for a single model.

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 (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://anonymous.4open.science/r/ksdagg-DBF7/README.md`.

Outline. Section 2 presents our framework and our notation. Section 3 introduces our algorithm KSDAGG (in Algorithm 1) and contains our main theoretical results. Section 4 presents numerical experiments to support KSDAGG. We close the paper with avenues for future research in Section 5.

2 Notation

We consider the goodness-of-fit problem where given access to a known probability density p (model) and to some i.i.d. d-dimensional samples $\mathbb{X}_N \coloneqq (X_i)_{i=1}^N$ 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 defined as the Maximum Mean Discrepancy (MMD, Gretton et al., 2012a) between p and q using the Stein kernel associated to k, that is

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

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

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

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

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

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 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, this is not true of the Stein kernel.

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]

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

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

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

For the *parametric bootstrap*, we directly draw new samples $(X'_i)_{i=1}^{N'}$ from the model p and compute the KSD

$$\bar{K}_k \coloneqq \frac{1}{N'(N'-1)} \sum_{1\leq i\neq j\leq N'} h_{p,k}(X'_i, X'_j). \tag{3}$$

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

$$\bar{K}_k \coloneqq \frac{1}{N(N-1)} \sum_{1\leq i\neq j\leq N} \epsilon_i \epsilon_j h_{p,k}(X_i, X_j). \tag{4}$$

[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.

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; Chwialkowski et al., 2014, 2016). 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* (Baraud, 2002).

Theorem 3.1. *Let $\psi := p-q$ and assume that $\max(\|p\|_\infty, \|q\|_\infty) \leq M$. Consider C_k as defined in Equation (2), $\alpha \in (0, e^{-1})$, $\beta \in (0,1)$ and $B_1 \in \mathbb{N}$ satisfying $B_1 \geq \frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$. 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 A, 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 $\psi := p-q$, 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 $C_k := \mathbb{E}_{q,q}[h_{p,k}(X,Y)^2]$. 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 (3) or (4), 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 $\big(\Delta^k_{u_\alpha w_k}\big)_{k\in\mathcal{K}}$ rejects it, that is

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

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

$$u_\alpha := \sup\left\{u \in \left(0, \min_{k\in\mathcal{K}} w_k^{-1}\right) : \widehat{P}_u \leq \alpha\right\} \tag{5}$$

where

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

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

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

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 (3) or (4)
  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 (3) or (4)
$u_{\min} = 0$, $u_{\max} = \min_{k\in\mathcal{K}} w_k^{-1}$
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: $\mathcal{O}\big(N^2(B_1+B_2)\,|\Lambda|\big)$

To compute u_α, we estimate the supremum in Equation (5) by performing B_3 steps of the bisection method. Detailed pseudocode for KSDAGG is provided in Algorithm 1.

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*

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

non-asymptotically using a parametric bootstrap and asymptotically using a wild bootstrap.

The proof of Proposition 3.2 is presented in Appendix B. We now provide guarantees for the power of our aggregated test KSDAGG in terms of its uniform separation rate.

Theorem 3.3. *Let $\psi := p - q$ denote the difference in densities and assume that $\|p\|_\infty \le M$ and $\|q\|_\infty \le M$. Consider the aggregated test $\Delta_\alpha^{\mathcal{K}}$ with a collection of kernels $\mathcal{K}$ and associated positive weights $(w_k)_{k\in\mathcal{K}}$ satisfying $\sum_{k\in\mathcal{K}} w_k \le 1$, and with parameters $\alpha \in (0, e^{-1})$ and $B_1, B_2, B_3 \in \mathbb{N}$ satisfying $B_1 \ge \frac{3}{\alpha^2}\big(\ln\big(\frac{8}{\beta}\big) + \alpha(1-\alpha)\big)$, $B_2 \ge \frac{8}{\alpha^2}\ln\big(\frac{2}{\beta}\big)$ and $B_3 \ge \log_2\big(\frac{4}{\alpha}\min_{k\in\mathcal{K}} w_k^{-1}\big)$. Consider some $\beta \in (0,1)$ and C_k as defined in Equation (2). 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, $\mathbb{P}_q\big(\Delta_\alpha^{\mathcal{K}}(\mathbb{X}_N) = 0\big) \le \beta$.

We prove Theorem 3.3 in Appendix C. 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

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

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

$$k_\lambda(x, y) = \left(1 + \frac{\|x - y\|_2^2}{\lambda^2}\right)^{-\beta_k} = \lambda^{2\beta_k}\Big(\lambda^2 + \|x - y\|_2^2\Big)^{-\beta_k} \propto \Big(\lambda^2 + \|x - y\|_2^2\Big)^{-\beta_k} \tag{6}$$

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. *Consider $\alpha \in (0, e^{-1})$, $\beta \in (0, 1)$ and $B_1, B_2, B_3 \in \mathbb{N}$ satisfying the conditions of Theorem 3.3 and assume that $\max(\|p\|_\infty, \|q\|_\infty) \leq M$. Given a collection Λ of positive bandwidths with associated positive weights $(w_\lambda)_{\lambda\in\Lambda}$ satisfying $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$, we consider $\mathcal{K}_\Lambda = \{k_\lambda : \lambda \in \Lambda\}$. There exists a positive constant C such that the condition*

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

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 $\mathcal{K}_\Lambda$. We do not impose any restrictions on $\psi := p - q$ such as assuming it belongs to a specific regularity class. For this reason, our result holds more generally but 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. For a particular regularity class, one can obtain a uniform separation rate N^{-r} for some $r > 0$ by choosing appropriate collections of bandwidths and weights (depending on N) such that the two terms have matching orders of N.

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

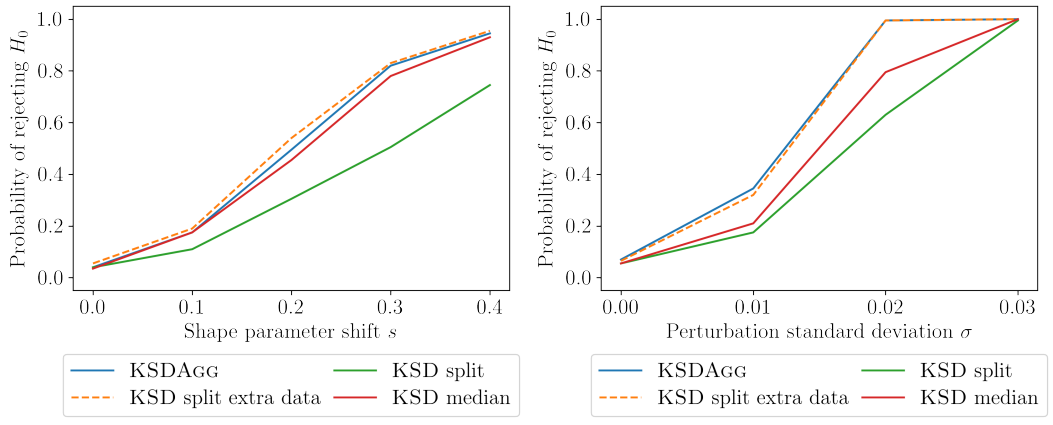
Gretton et al. (2012a) proposed to use the median heuristic as kernel bandwidth, it consists in the median of the L^2-distances between the samples given by

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

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 $\widehat{\text{KSD}}^2_{p,k}(\mathbb{X}_N) \,/\, \widehat{\sigma}_{\mathcal{H}_a}$ 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 (6) 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 (5). 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

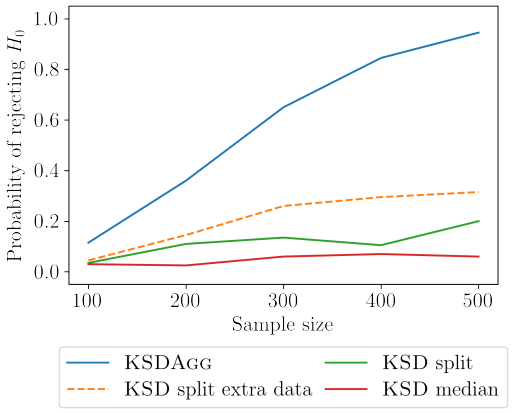
probability density p. As observed in Figure 2, some samples produced by the model can look exactly like MNIST digits, while other do not resemble digits. This Normalizing Flow has been trained to 'ideally' produce samples from the MNIST dataset. We are interested in whether or not we can detect the 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.

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 Discussion

We have introduced KSDAGG, an aggregated goodness-of-fit test based on the Kernel Stein Discrepancy. We have investigated the theoretical properties of this adaptive test. We have shown that it achieves the desired level and have provided conditions to guarantee high power by exhibiting a lower bound on its uniform separation rate. We have observed in our experiments that KSDAGG outperforms alternative state-of-the-art approaches to KSD kernel adaptation for goodness-of-fit testing.

This work covers 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.

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.

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).

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.

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.

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

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.

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.

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.

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.

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.

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

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

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

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

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

Liu, 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.

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*.

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

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*.

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*.

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 Section 5.
   (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://anonymous.4open.science/r/ksdagg-DBF7/README.md`.
   (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://anonymous.4open.science/r/ksdagg-DBF7/README.md`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]