A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, Arthur Gretton

A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum
Gatsby Unit, UCL
wittawatj@gmail.com

Wenkai Xu
Gatsby Unit, UCL
wenkaix@gatsby.ucl.ac.uk

Zoltán Szabó[*]
CMAP, École Polytechnique
zoltan.szabo@polytechnique.edu

Kenji Fukumizu
The Institute of Statistical Mathematics
fukumizu@ism.ac.jp

Arthur Gretton[*]
Gatsby Unit, UCL
arthur.gretton@gmail.com

Abstract

We propose a novel adaptive test of goodness-of-fit, with computational cost linear in the number of samples. We learn the test features that best indicate the differences between observed samples and a reference model, by minimizing the false negative rate. These features are constructed via Stein's method, meaning that it is not necessary to compute the normalising constant of the model. We analyse the asymptotic Bahadur efficiency of the new test, and prove that under a mean-shift alternative, our test always has greater relative efficiency than a previous linear-time kernel test, regardless of the choice of parameters for that test. In experiments, the performance of our method exceeds that of the earlier linear-time test, and matches or exceeds the power of a quadratic-time kernel test. In high dimensions and where model structure may be exploited, our goodness of fit test performs far better than a quadratic-time two-sample test based on the Maximum Mean Discrepancy, with samples drawn from the model.

1 Introduction

The goal of goodness of fit testing is to determine how well a model density $p(\mathbf{x})$ fits an observed sample $\mathsf{D} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ from an unknown distribution $q(\mathbf{x})$. This goal may be achieved via a hypothesis test, where the null hypothesis $H_0\colon p = q$ is tested against $H_1\colon p \neq q$. The problem of testing goodness of fit has a long history in statistics [11], with a number of tests proposed for particular parametric models. Such tests can require space partitioning [18, 3], which works poorly in high dimensions; or closed-form integrals under the model, which may be difficult to obtain, besides in certain special cases [2, 5, 30, 26]. An alternative is to conduct a two-sample test using samples drawn from *both* p and q. This approach was taken by [23], using a test based on the (quadratic-time) Maximum Mean Discrepancy [16], however this does not take advantage of the known structure of p (quite apart from the increased computational cost of dealing with samples from p).

More recently, measures of discrepancy with respect to a model have been proposed based on Stein's method [21]. A Stein operator for p may be applied to a class of test functions, yielding functions that have zero expectation under p. Classes of test functions can include the $W^{2,\infty}$ Sobolev space [14], and reproducing kernel Hilbert spaces (RKHS) [25]. Statistical tests have been proposed by [9, 22] based on classes of Stein transformed RKHS functions, where the test statistic is the norm of the smoothness-constrained function with largest expectation under q . We will refer to this statistic as the Kernel Stein Discrepancy (KSD). For consistent tests, it is sufficient to use C_0-universal kernels [6, Definition 4.1], as shown by [9, Theorem 2.2], although inverse multiquadric kernels may be preferred if uniform tightness is required [15].[2]

[*] Zoltán Szabó's ORCID ID: 0000-0001-6183-7603. Arthur Gretton's ORCID ID: 0000-0003-3169-7624.

[2] Briefly, [15] show that when an exponentiated quadratic kernel is used, a sequence of sets D may be constructed that does not correspond to any q, but for which the KSD nonetheless approaches zero. In a statistical testing setting, however, we assume identically distributed samples from q, and the issue does not arise.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

The minimum variance unbiased estimate of the KSD is a U-statistic, with computational cost quadratic in the number n of samples from q. It is desirable to reduce the cost of testing, however, so that larger sample sizes may be addressed. A first approach is to replace the U-statistic with a running average with linear cost, as proposed by [22] for the KSD, but this results in an increase in variance and corresponding decrease in test power. An alternative approach is to construct explicit features of the distributions, whose empirical expectations may be computed in linear time. In the two-sample and independence settings, these features were initially chosen at random by [10, 8, 32]. More recently, features have been constructed explicitly to maximize test power in the two-sample [19] and independence testing [20] settings, resulting in tests that are not only more interpretable, but which can yield performance matching quadratic-time tests.

We propose to construct explicit linear-time features for testing goodness of fit, chosen so as to maximize test power. These features further reveal where the model and data differ, in a readily interpretable way. Our first theoretical contribution is a derivation of the null and alternative distributions for tests based on such features, and a corresponding power optimization criterion. Note that the goodness-of-fit test requires somewhat different strategies to those employed for two-sample and independence testing [19, 20], which become computationally prohibitive in high dimensions for the Stein discrepancy (specifically, the normalization used in prior work to simplify the asymptotics would incur a cost cubic in the dimension d and the number of features in the optimization). Details may be found in Section 3.

Our second theoretical contribution, given in Section 4, is an analysis of the relative Bahadur efficiency of our test vs the linear time test of [22]: this represents the relative rate at which the p-value decreases under H_1 as we observe more samples. We prove that our test has greater asymptotic Bahadur efficiency relative to the test of [22], for Gaussian distributions under the mean-shift alternative. This is shown to hold regardless of the bandwidth of the exponentiated quadratic kernel used for the earlier test. The proof techniques developed are of independent interest, and we anticipate that they may provide a foundation for the analysis of relative efficiency of linear-time tests in the two-sample and independence testing domains. In experiments (Section 5), our new linear-time test is able to detect subtle local differences between the density $p(\mathbf{x})$, and the unknown $q(\mathbf{x})$ as observed through samples. We show that our linear-time test constructed based on optimized features has comparable performance to the quadratic-time test of [9, 22], while uniquely providing an explicit visual indication of where the model fails to fit the data.

2 Kernel Stein Discrepancy (KSD) Test

We begin by introducing the Kernel Stein Discrepancy (KSD) and associated statistical test, as proposed independently by [9] and [22]. Assume that the data domain is a connected open set $\mathcal{X} \subseteq \mathbb{R}^d$. Consider a Stein operator T_p that takes in a multivariate function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_d(\mathbf{x}))^\top \in \mathbb{R}^d$ and constructs a function $(T_p\mathbf{f})(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$. The constructed function has the key property that for all $\mathbf{f}$ in an appropriate function class, $\mathbb{E}_{\mathbf{x}\sim q}\left[(T_p\mathbf{f})(\mathbf{x})\right] = 0$ if and only if $q = p$. Thus, one can use this expectation as a statistic for testing goodness of fit.

The function class $\mathcal{F}^d$ for the function $\mathbf{f}$ is chosen to be a unit-norm ball in a reproducing kernel Hilbert space (RKHS) in [9, 22]. More precisely, let $\mathcal{F}$ be an RKHS associated with a positive definite kernel $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$. Let $\phi(\mathbf{x}) = k(\mathbf{x},\cdot)$ denote a feature map of k so that $k(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle_{\mathcal{F}}$. Assume that $f_i \in \mathcal{F}$ for all $i = 1,\ldots,d$ so that $\mathbf{f} \in \mathcal{F}\times\cdots\times\mathcal{F} := \mathcal{F}^d$ where $\mathcal{F}^d$ is equipped with the standard inner product $\langle\mathbf{f},\mathbf{g}\rangle_{\mathcal{F}^d} := \sum_{i=1}^d \langle f_i, g_i\rangle_{\mathcal{F}}$. The kernelized Stein operator T_p studied in [9] is $(T_p\mathbf{f})(\mathbf{x}) := \sum_{i=1}^d \left(\frac{\partial \log p(\mathbf{x})}{\partial x_i} f_i(\mathbf{x}) + \frac{\partial f_i(\mathbf{x})}{\partial x_i}\right) \overset{(a)}{=} \left\langle \mathbf{f}, \boldsymbol{\xi}_p(\mathbf{x},\cdot)\right\rangle_{\mathcal{F}^d}$, where at (a) we use the reproducing property of $\mathcal{F}$, i.e., $f_i(\mathbf{x}) = \langle f_i, k(\mathbf{x},\cdot)\rangle_{\mathcal{F}}$, and that $\frac{\partial k(\mathbf{x},\cdot)}{\partial x_i} \in \mathcal{F}$ [28, Lemma 4.34], hence $\boldsymbol{\xi}_p(\mathbf{x},\cdot) := \frac{\partial \log p(\mathbf{x})}{\partial \mathbf{x}} k(\mathbf{x},\cdot) + \frac{\partial k(\mathbf{x},\cdot)}{\partial \mathbf{x}}$ is in $\mathcal{F}^d$. We note that the Stein operator presented in [22] is defined such that $(T_p\mathbf{f})(\mathbf{x}) \in \mathbb{R}^d$. This distinction is not crucial and leads to the same goodness-of-fit test. Under appropriate conditions, e.g. that $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})f_i(\mathbf{x}) = 0$ for all $i = 1,\ldots,d$, it can be shown using integration by parts that $\mathbb{E}_{\mathbf{x}\sim p}(T_p\mathbf{f})(\mathbf{x}) = 0$ for any $\mathbf{f} \in \mathcal{F}^d$ [9, Lemma 5.1]. Based on the Stein operator, [9, 22] define the kernelized Stein discrepancy as

$$S_p(q) := \sup_{\|\mathbf{f}\|_{\mathcal{F}^d}\le 1} \mathbb{E}_{\mathbf{x}\sim q}\left\langle \mathbf{f}, \boldsymbol{\xi}_p(\mathbf{x},\cdot)\right\rangle_{\mathcal{F}^d} \overset{(a)}{=} \sup_{\|\mathbf{f}\|_{\mathcal{F}^d}\le 1} \left\langle \mathbf{f}, \mathbb{E}_{\mathbf{x}\sim q}\boldsymbol{\xi}_p(\mathbf{x},\cdot)\right\rangle_{\mathcal{F}^d} = \|\mathbf{g}(\cdot)\|_{\mathcal{F}^d}, \quad (1)$$

where at (a), $\boldsymbol{\xi}_p(\mathbf{x},\cdot)$ is Bochner integrable [28, Definition A.5.20] as long as $\mathbb{E}_{\mathbf{x}\sim q}\|\boldsymbol{\xi}_p(\mathbf{x},\cdot)\|_{\mathcal{F}^d} < \infty$, and $\mathbf{g}(\mathbf{y}) := \mathbb{E}_{\mathbf{x}\sim q}\boldsymbol{\xi}_p(\mathbf{x},\mathbf{y})$ is what we refer to as the *Stein witness function*. The Stein witness function will play a crucial role in our new test statistic in Section 3. When a C_0-universal kernel is used [6, Definition 4.1], and as long as $\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}}\log p(\mathbf{x}) - \nabla_{\mathbf{x}}\log q(\mathbf{x})\|^2 < \infty$, it can be shown that $S_p(q) = 0$ if and only if $p = q$ [9, Theorem 2.2].

The KSD $S_p(q)$ can be written as $S_p^2(q) = \mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}h_p(\mathbf{x},\mathbf{x}')$, where $h_p(\mathbf{x},\mathbf{y}) := \mathbf{s}_p^\top(\mathbf{x})\mathbf{s}_p(\mathbf{y})k(\mathbf{x},\mathbf{y}) + \mathbf{s}_p^\top(\mathbf{y})\nabla_{\mathbf{x}}k(\mathbf{x},\mathbf{y}) + \mathbf{s}_p^\top(\mathbf{x})\nabla_{\mathbf{y}}k(\mathbf{x},\mathbf{y}) + \sum_{i=1}^d \frac{\partial^2 k(\mathbf{x},\mathbf{y})}{\partial x_i \partial y_i}$, and $\mathbf{s}_p(\mathbf{x}) := \nabla_{\mathbf{x}}\log p(\mathbf{x})$ is a column vector. An unbiased empirical estimator of $S_p^2(q)$, denoted by $\widehat{S^2} = \frac{2}{n(n-1)}\sum_{i<j}h_p(\mathbf{x}_i,\mathbf{x}_j)$ [22, Eq. 14], is a degenerate U-statistic under H_0. For the goodness-of-fit test, the rejection threshold can be computed by a bootstrap procedure. All these properties make $\widehat{S^2}$ a very flexible criterion to detect the discrepancy of p and q: in particular, it can be computed even if p is known only up to a normalization constant. Further studies on nonparametric Stein operators can be found in [25, 14].

Linear-Time Kernel Stein (LKS) Test Computation of $\widehat{S^2}$ costs $\mathcal{O}(n^2)$. To reduce this cost, a linear-time (i.e., $\mathcal{O}(n)$) estimator based on an incomplete U-statistic is proposed in [22, Eq. 17], given by $\widehat{S_l^2} := \frac{2}{n}\sum_{i=1}^{n/2}h_p(\mathbf{x}_{2i-1},\mathbf{x}_{2i})$, where we assume n is even for simplicity. Empirically [22] observed that the linear-time estimator performs much worse (in terms of test power) than the quadratic-time U-statistic estimator, agreeing with our findings presented in Section 5.

3 New Statistic: The Finite Set Stein Discrepancy (FSSD)

Although shown to be powerful, the main drawback of the KSD test is its high computational cost of $\mathcal{O}(n^2)$. The LKS test is one order of magnitude faster. Unfortunately, the decrease in the test power outweighs the computational gain [22]. We therefore seek a variant of the KSD statistic that can be computed in linear time, and whose test power is comparable to the KSD test.

Key Idea The fact that $S_p(q) = 0$ if and only if $p = q$ implies that $\mathbf{g}(\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v} \in \mathcal{X}$ if and only if $p = q$, where $\mathbf{g}$ is the Stein witness function in (1). One can see $\mathbf{g}$ as a function witnessing the differences of p, q, in such a way that $|g_i(\mathbf{v})|$ is large when there is a discrepancy in the region around $\mathbf{v}$, as indicated by the i^{th} output of $\mathbf{g}$. The test statistic of [22, 9] is essentially given by the degree of "flatness" of $\mathbf{g}$ as measured by the RKHS norm $\|\cdot\|_{\mathcal{F}^d}$. The core of our proposal is to use a different measure of flatness of $\mathbf{g}$ which can be computed in linear time.

The idea is to use a real analytic kernel k which makes $g_1, \ldots, g_d$ real analytic. If $g_i \neq 0$ is an analytic function, then the Lebesgue measure of the set of roots $\{\mathbf{x} \mid g_i(\mathbf{x}) = 0\}$ is zero [24]. This property suggests that one can evaluate g_i at a finite set of locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$, drawn from a distribution with a density (w.r.t. the Lebesgue measure). If $g_i \neq 0$, then almost surely $g_i(\mathbf{v}_1), \ldots, g_i(\mathbf{v}_J)$ will not be zero. This idea was successfully exploited in recently proposed linear-time tests of [8] and [19, 20]. Our new test statistic based on this idea is called the Finite Set Stein Discrepancy (FSSD) and is given in Theorem 1. All proofs are given in the appendix.

Theorem 1 (The Finite Set Stein Discrepancy (FSSD))**.** *Let* $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} \subset \mathbb{R}^d$ *be random vectors drawn i.i.d. from a distribution* η *which has a density. Let* $\mathcal{X}$ *be a connected open set in* $\mathbb{R}^d$*. Define* $\mathrm{FSSD}_p^2(q) := \frac{1}{dJ}\sum_{i=1}^d\sum_{j=1}^J g_i^2(\mathbf{v}_j)$*. Assume that 1)* $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ *is* C_0*-universal [6, Definition 4.1] and real analytic i.e., for all* $\mathbf{v} \in \mathcal{X}$*,* $f(\mathbf{x}) := k(\mathbf{x},\mathbf{v})$ *is a real analytic function on* $\mathcal{X}$*. 2)* $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}h_p(\mathbf{x},\mathbf{x}') < \infty$*. 3)* $\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}}\log p(\mathbf{x}) - \nabla_{\mathbf{x}}\log q(\mathbf{x})\|^2 < \infty$*. 4)* $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = 0$.

Then, for any $J \geq 1$*,* η*-almost surely* $\mathrm{FSSD}_p^2(q) = 0$ *if and only if* $p = q$*.*

This measure depends on a set of J test locations (or features) $\{\mathbf{v}_i\}_{i=1}^J$ used to evaluate the Stein witness function, where J is fixed and is typically small. A kernel which is C_0-universal and real analytic is the Gaussian kernel $k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma_k^2}\right)$ (see [20, Proposition 3] for the result on analyticity). Throughout this work, we will assume all the conditions stated in Theorem 1, and consider only the Gaussian kernel. Besides the requirement that the kernel be real and analytic, the remaining conditions in Theorem 1 are the same as given in [9, Theorem 2.2]. Note that if the

FSSD is to be employed in a setting otherwise than testing, for instance to obtain pseudo-samples converging to p, then stronger conditions may be needed [15].

3.1 Goodness-of-Fit Test with the FSSD Statistic

Given a significance level α for the goodness-of-fit test, the test can be constructed so that H_0 is rejected when $n\widehat{\mathrm{FSSD}^2} > T_\alpha$, where T_α is the rejection threshold (critical value), and $\widehat{\mathrm{FSSD}^2}$ is an empirical estimate of $\mathrm{FSSD}_p^2(q)$. The threshold which guarantees that the type-I error (i.e., the probability of rejecting H_0 when it is true) is bounded above by α is given by the $(1-\alpha)$-quantile of the null distribution i.e., the distribution of $n\widehat{\mathrm{FSSD}^2}$ under H_0. In the following, we start by giving the expression for $\widehat{\mathrm{FSSD}^2}$, and summarize its asymptotic distributions in Proposition 2.

Let $\boldsymbol{\Xi}(\mathbf{x}) \in \mathbb{R}^{d\times J}$ such that $[\boldsymbol{\Xi}(\mathbf{x})]_{i,j} = \xi_{p,i}(\mathbf{x}, \mathbf{v}_j)/\sqrt{dJ}$. Define $\boldsymbol{\tau}(\mathbf{x}) := \mathrm{vec}(\boldsymbol{\Xi}(\mathbf{x})) \in \mathbb{R}^{dJ}$ where $\mathrm{vec}(\mathbf{M})$ concatenates columns of the matrix $\mathbf{M}$ into a column vector. We note that $\boldsymbol{\tau}(\mathbf{x})$ depends on the test locations $V = \{\mathbf{v}_j\}_{j=1}^J$. Let $\Delta(\mathbf{x},\mathbf{y}) := \boldsymbol{\tau}(\mathbf{x})^\top\boldsymbol{\tau}(\mathbf{y}) = \mathrm{tr}(\boldsymbol{\Xi}(\mathbf{x})^\top\boldsymbol{\Xi}(\mathbf{y}))$. Given an i.i.d. sample $\{\mathbf{x}_i\}_{i=1}^n \sim q$, a consistent, unbiased estimator of $\mathrm{FSSD}_p^2(q)$ is

$$\widehat{\mathrm{FSSD}^2} = \frac{1}{dJ}\sum_{l=1}^{d}\sum_{m=1}^{J}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\xi_{p,l}(\mathbf{x}_i,\mathbf{v}_m)\xi_{p,l}(\mathbf{x}_j,\mathbf{v}_m) = \frac{2}{n(n-1)}\sum_{i<j}\Delta(\mathbf{x}_i,\mathbf{x}_j), \quad (2)$$

which is a one-sample second-order U-statistic with Δ as its U-statistic kernel [27, Section 5.1.1]. Being a U-statistic, its asymptotic distribution can easily be derived. We use $\xrightarrow{d}$ to denote convergence in distribution.

Proposition 2 (Asymptotic distributions of $\widehat{\mathrm{FSSD}^2}$)**.** *Let $Z_1,\ldots,Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$, $\boldsymbol{\Sigma}_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p,q\}$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of $\boldsymbol{\Sigma}_p = \mathbb{E}_{\mathbf{x}\sim p}[\boldsymbol{\tau}(\mathbf{x})\boldsymbol{\tau}^\top(\mathbf{x})]$. Assume that $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{y}\sim q}\Delta^2(\mathbf{x},\mathbf{y}) < \infty$. Then, for any realization of $V = \{\mathbf{v}_j\}_{j=1}^J$, the following statements hold.*

1. *Under $H_0 : p = q$, $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2-1)\omega_i$.*
2. *Under $H_1 : p \neq q$, if $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_q\boldsymbol{\mu} > 0$, then $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0,\sigma_{H_1}^2)$.*

Proof. Recognizing that (2) is a degenerate U-statistic, the results follow directly from [27, Section 5.5.1, 5.5.2]. □

Claims 1 and 2 of Proposition 2 imply that under H_1, the test power (i.e., the probability of correctly rejecting H_1) goes to 1 asymptotically, if the threshold T_α is defined as above. In practice, simulating from the asymptotic null distribution in Claim 1 can be challenging, since the plug-in estimator of $\boldsymbol{\Sigma}_p$ requires a sample from p, which is not available. A straightforward solution is to draw sample from p, either by assuming that p can be sampled easily or by using a Markov chain Monte Carlo (MCMC) method, although this adds an additional computational burden to the test procedure. A more subtle issue is that when dependent samples from p are used in obtaining the test threshold, the test may become more conservative than required for i.i.d. data [7]. An alternative approach is to use the plug-in estimate $\hat{\boldsymbol{\Sigma}}_q$ instead of $\boldsymbol{\Sigma}_p$. The covariance matrix $\hat{\boldsymbol{\Sigma}}_q$ can be directly computed from the data. This is the approach we take. Theorem 3 guarantees that the replacement of the covariance in the computation of the asymptotic null distribution still yields a consistent test. We write $\mathbb{P}_{H_1}$ for the distribution of $n\widehat{\mathrm{FSSD}^2}$ under H_1.

Theorem 3. *Let $\hat{\boldsymbol{\Sigma}}_q := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}(\mathbf{x}_i)\boldsymbol{\tau}^\top(\mathbf{x}_i) - [\frac{1}{n}\sum_{i=1}^n\boldsymbol{\tau}(\mathbf{x}_i)][\frac{1}{n}\sum_{j=1}^n\boldsymbol{\tau}(\mathbf{x}_j)]^\top$ with $\{\mathbf{x}_i\}_{i=1}^n \sim q$. Suppose that the test threshold T_α is set to the $(1-\alpha)$-quantile of the distribution of $\sum_{i=1}^{dJ}(Z_i^2-1)\hat{\nu}_i$ where $\{Z_i\}_{i=1}^{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $\hat{\nu}_1,\ldots,\hat{\nu}_{dJ}$ are eigenvalues of $\hat{\boldsymbol{\Sigma}}_q$. Then, under H_0, asymptotically the false positive rate is α. Under H_1, for $\{\mathbf{v}_j\}_{j=1}^J$ drawn from a distribution with a density, the test power $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha) \to 1$ as $n \to \infty$.*

Remark 1. The proof of Theorem 3 relies on two facts. First, under H_0, $\hat{\boldsymbol{\Sigma}}_q = \hat{\boldsymbol{\Sigma}}_p$ i.e., the plug-in estimate of $\boldsymbol{\Sigma}_p$. Thus, under H_0, the null distribution approximated with $\hat{\boldsymbol{\Sigma}}_q$ is asymptotically correct, following the convergence of $\hat{\mathbf{\Sigma}}_p$ to $\mathbf{\Sigma}_p$. Second, the rejection threshold obtained from the approximated null distribution is asymptotically constant. Hence, under H_1, claim 2 of Proposition 2 implies that $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \infty$ as $n \to \infty$, and consequently $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha) \to 1$.

3.2 Optimizing the Test Parameters

Theorem 1 guarantees that the population quantity $\mathrm{FSSD}^2 = 0$ if and only if $p = q$ for any choice of $\{\mathbf{v}_i\}_{i=1}^J$ drawn from a distribution with a density. In practice, we are forced to rely on the empirical $\widehat{\mathrm{FSSD}^2}$, and some test locations will give a higher detection rate (i.e., test power) than others for finite n. Following the approaches of [17, 20, 19, 29], we choose the test locations $V = \{\mathbf{v}_j\}_{j=1}^J$ and kernel bandwidth σ_k^2 so as to maximize the test power i.e., the probability of rejecting H_0 when it is false. We first give an approximate expression for the test power when n is large.

Proposition 4 (Approximate test power of $n\widehat{\mathrm{FSSD}^2}$)**.** *Under H_1, for large n and fixed r, the test power $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}\right)$, where Φ denotes the cumulative distribution function of the standard normal distribution, and σ_{H_1} is defined in Proposition 2.*

Proof. $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > r) = \mathbb{P}_{H_1}(\widehat{\mathrm{FSSD}^2} > r/n) = \mathbb{P}_{H_1}\left(\sqrt{n}\frac{\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2}{\sigma_{H_1}} > \sqrt{n}\frac{r/n - \mathrm{FSSD}^2}{\sigma_{H_1}}\right)$. For sufficiently large n, the alternative distribution is approximately normal as given in Proposition 2. It follows that $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}\right)$. □

Let $\boldsymbol{\zeta} := \{V, \sigma_k^2\}$ be the collection of all tuning parameters. Assume that n is sufficiently large. Following the same argument as in [29], in $\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}$, we observe that the first term $\frac{r}{\sqrt{n}\sigma_{H_1}} = \mathcal{O}(n^{-1/2})$ going to 0 as $n \to \infty$, while the second term $\sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}} = \mathcal{O}(n^{1/2})$, dominating the first for large n. Thus, the best parameters that maximize the test power are given by $\boldsymbol{\zeta}^* = \arg\max_{\boldsymbol{\zeta}} \mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha) \approx \arg\max_{\boldsymbol{\zeta}} \frac{\mathrm{FSSD}^2}{\sigma_{H_1}}$. Since FSSD^2 and σ_{H_1} are unknown, we divide the sample $\{\mathbf{x}_i\}_{i=1}^n$ into two disjoint training and test sets, and use the training set to compute $\frac{\widehat{\mathrm{FSSD}^2}}{\hat{\sigma}_{H_1} + \gamma}$, where a small regularization parameter $\gamma > 0$ is added for numerical stability. The goodness-of-fit test is performed on the test set to avoid overfitting. The idea of splitting the data into training and test sets to learn good features for hypothesis testing was successfully used in [29, 20, 19, 17].

To find a local maximum of $\frac{\widehat{\mathrm{FSSD}^2}}{\hat{\sigma}_{H_1} + \gamma}$, we use gradient ascent for its simplicity. The initial points of $\{\mathbf{v}_i\}_{i=1}^J$ are set to random draws from a normal distribution fitted to the training data, a heuristic we found to perform well in practice. The objective is non-convex in general, reflecting many possible ways to capture the differences of p and q. The regularization parameter γ is not tuned, and is fixed to a small constant. Assume that $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ costs $\mathcal{O}(d^2)$ to evaluate. Computing $\nabla_{\boldsymbol{\zeta}} \frac{\widehat{\mathrm{FSSD}^2}}{\hat{\sigma}_{H_1} + \gamma}$ costs $\mathcal{O}(d^2 J^2 n)$. The computational complexity of $n\widehat{\mathrm{FSSD}^2}$ and $\hat{\sigma}_{H_1}^2$ is $\mathcal{O}(d^2 J n)$. Thus, finding a local optimum via gradient ascent is still linear-time, for a fixed maximum number of iterations. Computing $\hat{\mathbf{\Sigma}}_q$ costs $\mathcal{O}(d^2 J^2 n)$, and obtaining all the eigenvalues of $\hat{\mathbf{\Sigma}}_q$ costs $\mathcal{O}(d^3 J^3)$ (required only once). If the eigenvalues decay to zero sufficiently rapidly, one can approximate the asymptotic null distribution with only a few eigenvalues. The cost to obtain the largest few eigenvalues alone can be much smaller.

Remark 2. Let $\hat{\boldsymbol{\mu}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}(\mathbf{x}_i)$. It is possible to normalize the FSSD statistic to get a new statistic $\hat{\lambda}_n := n\hat{\boldsymbol{\mu}}^\top(\hat{\mathbf{\Sigma}}_q + \gamma\mathbf{I})^{-1}\hat{\boldsymbol{\mu}}$ where $\gamma \geq 0$ is a regularization parameter that goes to 0 as $n \to \infty$. This was done in the case of the ME (mean embeddings) statistic of [8, 19]. The asymptotic null distribution of this statistic takes the convenient form of $\chi^2(dJ)$ (independent of p and q), eliminating the need to obtain the eigenvalues of $\hat{\mathbf{\Sigma}}_q$. It turns out that the test power criterion for tuning the parameters in this case is the statistic $\hat{\lambda}_n$ itself. However, the optimization is computationally expensive as $(\hat{\mathbf{\Sigma}}_q + \gamma\mathbf{I})^{-1}$ (costing $\mathcal{O}(d^3 J^3)$) needs to be reevaluated in each gradient ascent iteration. This is not needed in our proposed FSSD statistic.

4 Relative Efficiency and Bahadur Slope

Both the linear-time kernel Stein (LKS) and FSSD tests have the same computational cost of $\mathcal{O}(d^2 n)$, and are consistent, achieving maximum power of 1 as $n \to \infty$ under H_1. It is thus of theoretical interest to understand which test is more sensitive in detecting the differences of p and q. This can be quantified by the *Bahadur slope* of the test [1]. Two given tests can then be compared by computing the *Bahadur efficiency* (Theorem 7) which is given by the ratio of the slopes of the two tests. We note that the constructions and techniques in this section may be of independent interest, and can be generalised to other statistical testing settings.

We start by introducing the concept of Bahadur slope for a general test, following the presentation of [12, 13]. Consider a hypothesis testing problem on a parameter θ. The test proposes a null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative hypothesis $H_1 : \theta \in \Theta \backslash \Theta_0$, where Θ, Θ_0 are arbitrary sets. Let T_n be a test statistic computed from a sample of size n, such that large values of T_n provide an evidence to reject H_0. We use plim to denote convergence in probability, and write $\mathbb{E}_r$ for $\mathbb{E}_{\mathbf{x}\sim r}\mathbb{E}_{\mathbf{x}'\sim r}$.

Approximate Bahadur Slope (ABS) For $\theta_0 \in \Theta_0$, let the asymptotic null distribution of T_n be $F(t) = \lim_{n\to\infty} P_{\theta_0}(T_n < t)$, where we assume that the CDF (F) is continuous and common to all $\theta_0 \in \Theta_0$. The continuity of F will be important later when Theorem 9 and 10 are used to compute the slopes of LKS and FSSD tests. Assume that there exists a continuous strictly increasing function $\rho : (0, \infty) \to (0, \infty)$ such that $\lim_{n\to\infty} \rho(n) = \infty$, and that $-2 \operatorname{plim}_{n\to\infty} \frac{\log(1-F(T_n))}{\rho(n)} = c(\theta)$ where $T_n \sim P_\theta$, for some function c such that $0 < c(\theta_A) < \infty$ for $\theta_A \in \Theta\backslash\Theta_0$, and $c(\theta_0) = 0$ when $\theta_0 \in \Theta_0$. The function $c(\theta)$ is known as the *approximate Bahadur slope* (ABS) of the sequence T_n. The quantifier "approximate" comes from the use of the asymptotic null distribution instead of the exact one [1]. Intuitively the slope $c(\theta_A)$, for $\theta_A \in \Theta\backslash\Theta_0$, is the rate of convergence of p-values (i.e., $1 - F(T_n)$) to 0, as n increases. The higher the slope, the faster the p-value vanishes, and thus the lower the sample size required to reject H_0 under θ_A.

Approximate Bahadur Efficiency Given two sequences of test statistics, $T_n^{(1)}$ and $T_n^{(2)}$ having the same $\rho(n)$ (see Theorem 10), the approximate Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$ is defined as $E(\theta_A) := c^{(1)}(\theta_A)/c^{(2)}(\theta_A)$ for $\theta_A \in \Theta\backslash\Theta_0$. If $E(\theta_A) > 1$, then $T_n^{(1)}$ is asymptotically more efficient than $T_n^{(2)}$ in the sense of Bahadur, for the particular problem specified by $\theta_A \in \Theta\backslash\Theta_0$. We now give approximate Bahadur slopes for two sequences of linear time test statistics: the proposed $n\widehat{\mathrm{FSSD}^2}$, and the LKS test statistic $\sqrt{n}\widehat{S_l^2}$ discussed in Section 2.

Theorem 5. *The approximate Bahadur slope of* $n\widehat{\mathrm{FSSD}^2}$ *is* $c^{(\mathrm{FSSD})} := \mathrm{FSSD}^2/\omega_1$, *where* ω_1 *is the maximum eigenvalue of* $\mathbf{\Sigma}_p := \mathbb{E}_{\mathbf{x}\sim p}[\boldsymbol{\tau}(\mathbf{x})\boldsymbol{\tau}^\top(\mathbf{x})]$ *and* $\rho(n) = n$.

Theorem 6. *The approximate Bahadur slope of the linear-time kernel Stein (LKS) test statistic* $\sqrt{n}\widehat{S_l^2}$ *is* $c^{(\mathrm{LKS})} = \frac{1}{2}\frac{\left[\mathbb{E}_q h_p(\mathbf{x},\mathbf{x}')\right]^2}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$, *where* h_p *is the U-statistic kernel of the KSD statistic, and* $\rho(n) = n$.

To make these results concrete, we consider the setting where $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, 1)$. We assume that both tests use the Gaussian kernel $k(x, y) = \exp\left(-(x-y)^2/2\sigma_k^2\right)$, possibly with different bandwidths. We write σ_k^2 and κ^2 for the FSSD and LKS bandwidths, respectively. Under these assumptions, the slopes given in Theorem 5 and Theorem 6 can be derived explicitly. The full expressions of the slopes are given in Proposition 12 and Proposition 13 (in the appendix). By [12, 13] (recalled as Theorem 10 in the supplement), the approximate Bahadur efficiency can be computed by taking the ratio of the two slopes. The efficiency is given in Theorem 7.

Theorem 7 (Efficiency in the Gaussian mean shift problem)**.** *Let* $E_1(\mu_q, v, \sigma_k^2, \kappa^2)$ *be the approximate Bahadur efficiency of* $n\widehat{\mathrm{FSSD}^2}$ *relative to* $\sqrt{n}\widehat{S_l^2}$ *for the case where* $p = \mathcal{N}(0, 1), q = \mathcal{N}(\mu_q, 1)$, *and* $J = 1$ *(i.e., one test location* v *for* $n\widehat{\mathrm{FSSD}^2}$*). Fix* $\sigma_k^2 = 1$ *for* $n\widehat{\mathrm{FSSD}^2}$. *Then, for any* $\mu_q \neq 0$, *for some* $v \in \mathbb{R}$, *and for any* $\kappa^2 > 0$, *we have* $E_1(\mu_q, v, \sigma_k^2, \kappa^2) > 2$.

When $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, 1)$ for $\mu_q \neq 0$, Theorem 7 guarantees that our FSSD test is asymptotically at least twice as efficient as the LKS test in the Bahadur sense. We note that the efficiency is conservative in the sense that $\sigma_k^2 = 1$ regardless of μ_q. Choosing σ_k^2 dependent on μ_q will likely improve the efficiency further.

5 Experiments

In this section, we demonstrate the performance of the proposed test on a number of problems. The primary goal is to understand the conditions under which the test can perform well.

Sensitivity to Local Differences We start by demonstrating that the test power objective $\mathrm{FSSD}^2/\sigma_{H_1}$ captures local differences of p and q, and that interpretable features v are found. Consider a one-dimensional problem in which $p = \mathcal{N}(0, 1)$ and $q = \mathrm{Laplace}(0, 1/\sqrt{2})$, a zero-mean Laplace distribution with scale parameter $1/\sqrt{2}$. These parameters are chosen so that p and q have the same mean and variance. Figure 1 plots the (rescaled) objective as a function of v. The objective illustrates that the best features (indicated by v^*) are at the most discriminative locations.

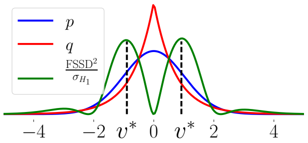


Figure 1: The power criterion $\mathrm{FSSD}^2/\sigma_{H_1}$ as a function of test location v.

Test Power We next investigate the power of different tests on two problems:

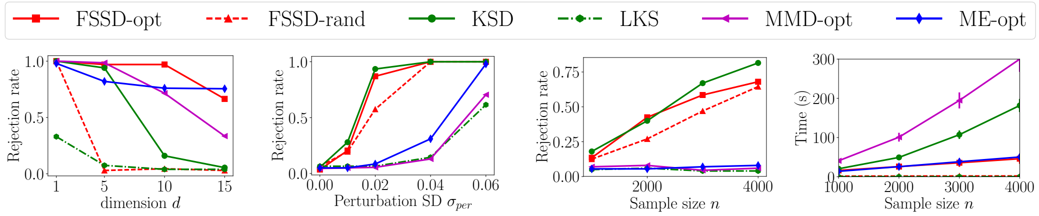
1. **Gaussian vs. Laplace**: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_d)$ and $q(\mathbf{x}) = \prod_{i=1}^{d} \mathrm{Laplace}(x_i|0, 1/\sqrt{2})$ where the dimension d will be varied. The two distributions have the same mean and variance. The main characteristic of this problem is local differences of p and q (see Figure 1). Set $n = 1000$.

2. **Restricted Boltzmann Machine** (RBM): $p(\mathbf{x})$ is the marginal distribution of $p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right)$, where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{h} \in \{\pm 1\}^{d_h}$ is a random vector of hidden variables, and Z is the normalization constant. The exact marginal density $p(\mathbf{x}) = \sum_{\mathbf{h} \in \{-1,1\}^{d_h}} p(\mathbf{x}, \mathbf{h})$ is intractable when d_h is large, since it involves summing over 2^{d_h} terms. Recall that the proposed test only requires the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ (not the normalization constant), which can be computed in closed form in this case. In this problem, q is another RBM where entries of the matrix $\mathbf{B}$ are corrupted by Gaussian noise. This was the problem considered in [22]. We set $d = 50$ and $d_h = 40$, and generate samples by n independent chains (i.e., n independent samples) of blocked Gibbs sampling with 2000 burn-in iterations.

We evaluate the following six kernel-based nonparametric tests with $\alpha = 0.05$, all using the Gaussian kernel. **1. FSSD-rand**: the proposed FSSD test where the test locations set to random draws from a multivariate normal distribution fitted to the data. The kernel bandwidth is set by the commonly used median heuristic i.e., $\sigma_k = \mathrm{median}(\{\|\mathbf{x}_i - \mathbf{x}_j\|, i < j\})$. **2. FSSD-opt**: the proposed FSSD test where both the test locations and the Gaussian bandwidth are optimized (Section 3.2). **3. KSD**: the quadratic-time Kernel Stein Discrepancy test with the median heuristic. **4. LKS**: the linear-time version of KSD with the median heuristic. **5. MMD-opt**: the quadratic-time MMD two-sample test of [16] where the kernel bandwidth is optimized by grid search to maximize a power criterion as described in [29]. **6. ME-opt**: the linear-time mean embeddings (ME) two-sample test of [19] where parameters are optimized. We draw n samples from p to run the two-sample tests (MMD-opt, ME-opt). For FSSD tests, we use $J = 5$ (see Section A for an investigation of test power as J varies). All tests with optimization use 20% of the sample size n for parameter tuning. Code is available at `https://github.com/wittawatj/kernel-gof`.

Figure 2 shows the rejection rates of the six tests for the two problems, where each problem is repeated for 200 trials, resampling n points from q every time. In Figure 2a (Gaussian vs. Laplace), high performance of FSSD-opt indicates that the test performs well when there are local differences between p and q. Low performance of FSSD-rand emphasizes the importance of the optimization of FSSD-opt to pinpoint regions where p and q differ. The power of KSD quickly drops as the dimension increases, which can be understood since KSD is the RKHS norm of a function witnessing differences in p and q across the entire domain, including where these differences are small.

We next consider the case of RBMs. Following [22], $\mathbf{b}, \mathbf{c}$ are independently drawn from the standard multivariate normal distribution, and entries of $\mathbf{B} \in \mathbb{R}^{50 \times 40}$ are drawn with equal probability from $\{\pm 1\}$, in each trial. The density q represents another RBM having the same $\mathbf{b}, \mathbf{c}$ as in p, and with all entries of $\mathbf{B}$ corrupted by independent zero-mean Gaussian noise with standard deviation σ_{per}. Figure



(a) Gaussian vs. Laplace. $n = 1000$.

(b) RBM. $n = 1000$. Perturb all entries of $\mathbf{B}$.

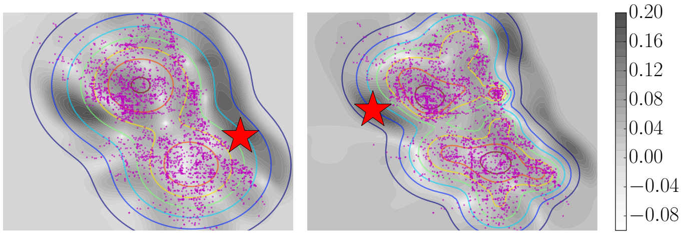
(c) RBM. $\sigma_{per} = 0.1$. Perturb $B_{1,1}$.

(d) Runtime (RBM)

Figure 2: Rejection rates of the six tests. The proposed linear-time FSSD-opt has a comparable or higher test power in some cases than the quadratic-time KSD test.

2b shows the test powers as σ_{per} increases, for a fixed sample size $n = 1000$. We observe that all the tests have correct false positive rates (type-I errors) at roughly $\alpha = 0.05$ when there is no perturbation noise. In particular, the optimization in FSSD-opt does not increase false positive rate when H_0 holds. We see that the performance of the proposed FSSD-opt matches that of the quadratic-time KSD at all noise levels. MMD-opt and ME-opt perform far worse than the goodness-of-fit tests when the difference in p and q is small (σ_{per} is low), since these tests simply represent p using samples, and do not take advantage of its structure.

The advantage of having $\mathcal{O}(n)$ runtime can be clearly seen when the problem is much harder, requiring larger sample sizes to tackle. Consider a similar problem on RBMs in which the parameter $\mathbf{B} \in \mathbb{R}^{50\times 40}$ in q is given by that of p, where only the first entry $B_{1,1}$ is perturbed by random $\mathcal{N}(0, 0.1^2)$ noise. The results are shown in Figure 2c where the sample size n is varied. We observe that the two two-sample tests fail to detect this subtle difference even with large sample size. The test powers of KSD and FSSD-opt are comparable when n is relatively small. It appears that KSD has higher test power than FSSD-opt in this case for large n. However, this moderate gain in the test power comes with an order of magnitude more computation. As shown in Figure 2d, the runtime of the KSD is much larger than that of FSSD-opt, especially at large n. In these problems, the performance of the new test (even without optimization) far exceeds that of the LKS test. Further simulation results can be found in Section B.



(a) p = 2-component GMM.

(b) p = 10-component GMM

Figure 3: Plots of the optimization objective as a function of test location $\mathbf{v} \in \mathbb{R}^2$ in the Gaussian mixture model (GMM) evaluation task.

Interpretable Features In the final simulation, we demonstrate that the learned test locations are informative in visualising where the model does not fit the data well. We consider crime data from the Chicago Police Department, recording $n = 11957$ locations (latitude-longitude coordinates) of robbery events in Chicago in 2016.[3] We address the situation in which a model p for the robbery location density is given, and we wish to visualise where it fails to match the data. We fit a Gaussian mixture model (GMM) with the expectation-maximization algorithm to a subsample of 5500 points. We then test the model on a held-out test set of the same size to obtain proposed locations of relevant features $\mathbf{v}$. Figure 3a shows the test robbery locations in purple, the model with two Gaussian components in wireframe, and the optimization objective for $\mathbf{v}$ as a grayscale contour plot (a red star indicates the maximum). We observe that the 2-component model is a poor fit to the data, particularly in the right tail areas of the data, as indicated in dark gray (i.e., the objective is high). Figure 3b shows a similar plot with a 10-component GMM. The additional components appear to have eliminated some mismatch in the right tail, however a discrepancy still exists in the left region. Here, the data have a sharp boundary on the right side following the geography of Chicago, and do not exhibit exponentially decaying Gaussian-like tails. We note that tests based on a learned feature located at the maximum both correctly reject H_0.

[3] Data can be found at https://data.cityofchicago.org.

Acknowledgement

WJ, WX, and AG thank the Gatsby Charitable Foundation for the financial support. ZSz was financially supported by the Data Science Initiative. KF has been supported by KAKENHI Innovative Areas 25120012.

References

[1] R. R. Bahadur. Stochastic comparison of tests. *The Annals of Mathematical Statistics*, 31(2): 276–295, 1960.

[2] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988.

[3] J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318, 1994.

[4] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[5] A. Bowman and P. Foster. Adaptive smoothing and density based tests of multivariate normality. *Journal of the American Statistical Association*, 88:529–537, 1993.

[6] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, Jan. 2010.

[7] K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *NIPS*, pages 3608–3616, 2014.

[8] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pages 1981–1989, 2015.

[9] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *ICML*, pages 2606–2615, 2016.

[10] T. Epps and K. Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3–4):177–203, 1986.

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

[12] L. J. Gleser. On a measure of test efficiency proposed by R. R. Bahadur. 35(4):1537–1544, 1964.

[13] L. J. Gleser. The comparison of multivariate tests of hypothesis by means of Bahadur efficiency. 28(2):157–174, 1966.

[14] J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *NIPS*, pages 226–234, 2015.

[15] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *ICML*, pages 1292–1301. PMLR, 06–11 Aug 2017.

[16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

[17] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, pages 1205–1213. 2012.

[18] L. Györfi and E. C. van der Meulen. A consistent goodness of fit test based on the total variation distance. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 631–645, 1990.

[19] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable Distribution Features with Maximum Testing Power. In *NIPS*, pages 181–189. 2016.

[20] W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In *ICML*, pages 1742–1751. PMLR, 2017.

[21] C. Ley, G. Reinert, and Y. Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017.

[22] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, pages 276–284, 2016.

[23] J. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *NIPS*, pages 829–837, 2015.

[24] B. Mityagin. The Zero Set of a Real Analytic Function. Dec. 2015. arXiv: 1512.07276.

[25] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

[26] M. L. Rizzo. New goodness-of-fit tests for Pareto distributions. *ASTIN Bulletin: Journal of the International Association of Actuaries*, 39(2):691–715, 2009.

[27] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

[28] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

[29] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized Maximum Mean Discrepancy. In *ICLR*, 2016.

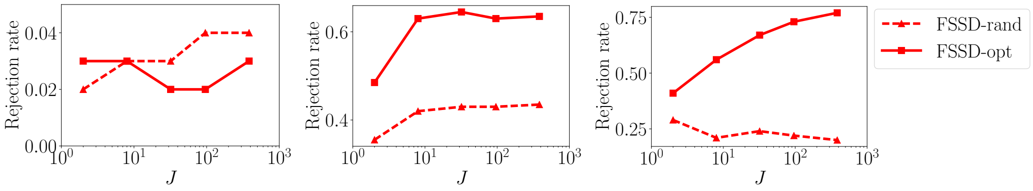
[30] G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

[31] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[32] Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18, 2017.

A Linear-Time Kernel Goodness-of-Fit Test Supplementary

A Rejection Rate vs. Number of Test Locations J



(a) SG. $d = 5$. $\alpha = 0.05$ (b) Gaussian vs. GMM. $d = 1$. (c) GVD. $d = 5$.

Figure 4: Plots of rejection rate against the number of test locations J in the three toy problems in Section A.

The aim of this section is to explore the test power of the proposed FSSD test as a function of the number of test locations J. We consider three synthetic problems to illustrate three phenomena depending on the characteristic of the problem. We note that the test power may not necessarily increase with J. Figure 4 shows the rejection rate as a function of the test locations J in the three problems described below. In all cases, the sample size is set to $n = 500$, the train/test ratio is 50%, and the significance level is $\alpha = 0.05$. All rejection rates are computed with 200 trials with data sampled from the specified q in every trial.

We emphasize that the FSSD test is not designed to be used with large J, since doing so defeats the purpose of a linear-time test. We show in the main text in Section 2 that using $J = 5$ is typically sufficient in practice.

Same Gaussian (SG): In this problem, $p = q = \mathcal{N}(\mathbf{0}, \mathbf{I})$ in $\mathbb{R}^5$ i.e., H_0 is true. It can be seen in Figure 4a that both the FSSD tests with and without optimization achieve correct false positive rate at roughly α for all J considered. That is, under H_0, the false rejection rate stays at the right level for all J.

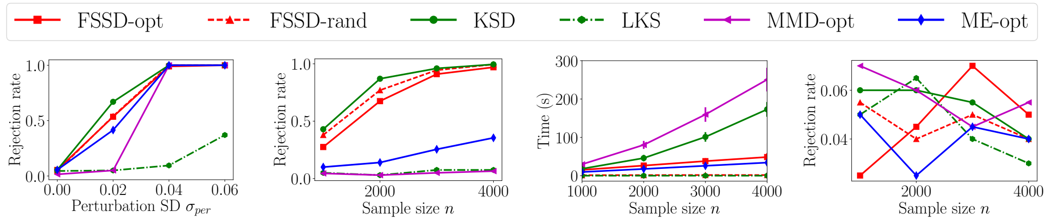
Gaussian vs. Gaussian mixture model (GMM): This is a one-dimensional problem where $p = \mathcal{N}(0, 1)$ and $q = 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 0.1^2)$ i.e., a mixture of two normal distributions. In this problem, p significantly differs from q in a small region around 0. This difference is created by the second mixture component. The characteristic of this problem is the local difference of p and q.

Figure 4b indicates that using random test locations (FSSD-rand) does not give high test power. With optimization (FSSD-opt), the power increases as J increases up to a point, after which it slightly drops down and reaches a plateau. This behavior can be explained by noting that there is only a very small region around 0 to detect the difference. More signal can be gained with diminishing return by increasing the number of test locations around 0. When J is sufficiently high, the increase in the variance of the statistic outweighs the gain of the signal (recall that the variance of the null distribution increases with J). This increase in the variance reduces the test power.

Gaussian Variance Difference (GVD): This is a synthetic problem studied in [19] where $p = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $q = \mathcal{N}(\mathbf{0}, \mathrm{diag}(2, 1 \ldots, 1))$ in $\mathbb{R}^5$. In this case, the region of difference between q and p exists only along the first dimension, and is broad.

In this case, Figure 4c shows that, with optimization, the power increases as the number of test locations increases. Unlike the case of Gaussian vs. GMM, the region of difference in this case is broad, and can accommodate more test locations to increase the signal. Despite this, we expect the test power to reach a plateau when J is sufficiently large for the same reason as described previously. In FSSD-rand, random test locations decrease the power due to the increase in the variance. Since only one dimension is relevant in determining the difference of p and q, it is unlikely that random locations are in the right region.

B More Experiments



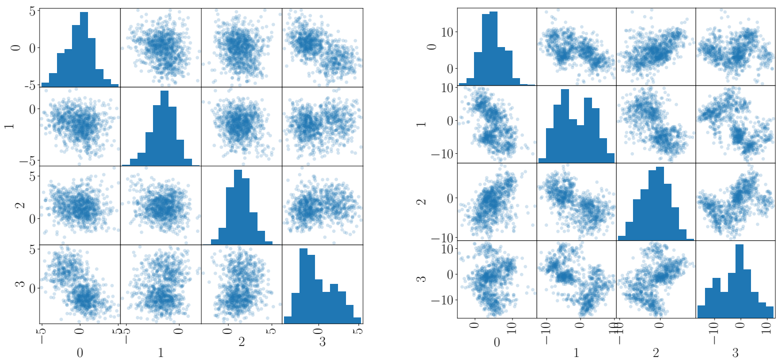
(a) RBM. $n = 1000$. Perturb all entries of $\mathbf{B}$.

(b) RBM. $\sigma_{per} = 0.1$. Perturb $B_{1,1}$.

(c) Runtime (RBM)

(d) RBM. No perturbation. H_0 holds.

Figure 5: Rejection rates of the six tests in the RBM problem with $d = 50$ and $d_h = 10$.



(a) $d = 50, d_h = 10$

(b) $d = 50, d_h = 40$

Figure 6: Pairwise scatter plots of 1000 points drawn from RBMs. Only the first 4 variates out of 50 are shown. **(a)**: RBM with $d = 50$ dimensions with $d_h = 10$ latent variables. **(b)**: RBM with $d = 50$ dimensions with $d_h = 40$ latent variables.

Recall that in Section 5, we evaluate the test powers of all the six tests on the RBM problem with $d = 50$ and $d_h = 40$ (i.e., the number of latent variables). We aim to provide more evaluations in this section. In [22], the setting of $d = 50$ and $d_h = 10$ was studied. Here we consider the same setting and show the results in Figure 5 where all other problem configurations are the same as in Section 5.

In Figure 5a, p is set to an RBM with parameters randomly drawn (described in Section 5), and q is the same RBM with all entries of the parameter $\mathbf{B} \in \mathbb{R}^{50\times 10}$ perturbed by independent Gaussian noise with standard deviation σ_{per}, which varies from 0 to 0.06. We observe that the proposed FSSD-opt and KSD perform comparably. Figure 5b considers a hard problem where only the first entry $B_{1,1}$ is perturbed by noise following $\mathcal{N}(0, 0.1^2)$, and the sample size n is varied. In both of these two cases, the overall trend is similar to the case of $d = 50$ and $d_h = 40$ presented in Figure 2. It is interesting to note that FSSD-rand, relying on random test locations, performs comparably or even outperforms FSSD-opt in the case of $d = 50, d_h = 10$, but not in the case of $d = 50, d_h = 40$. This phenomenon can be explained as follows. In the case of $d = 50, d_h = 10$, the data generated from the RBM tend to have simple structure (see Figure 6a). By contrast, data generated from the RBM with $d = 50, d_h = 40$ (more latent variables) have larger variance, and can form a complicated structure (Figure 6b), requiring a careful choice of test locations to detect differences of p and q. When $d = 50, d_h = 10$, however, random test locations given by random draws from a Gaussian distribution fitted to the data are sufficient to capture the simple structural difference. This explains why FSSD-rand can perform well in this case. Additionally, FSSD-rand also has 20% more testing data, since FSSD-opt uses 20% of the sample for parameter tuning.

Figure 5d shows the rejection rates of all the tests as the sample size increases when p and q are the same RBM. All the tests have roughly the right false rejection rates at the set significance level $\alpha = 0.05$.

C Proof of Theorem 1

Recall Theorem 1:

Theorem 1 (The Finite Set Stein Discrepancy (FSSD)). *Let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} \subset \mathbb{R}^d$ be random vectors drawn i.i.d. from a distribution η which has a density. Let $\mathcal{X}$ be a connected open set in $\mathbb{R}^d$. Define $\mathrm{FSSD}^2_p(q) := \frac{1}{dJ}\sum_{i=1}^{d}\sum_{j=1}^{J} g_i^2(\mathbf{v}_j)$. Assume that $\underline{1)}$ $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is C_0-universal [6, Definition 4.1] and real analytic i.e., for all $\mathbf{v} \in \mathcal{X}$, $f(\mathbf{x}) := k(\mathbf{x}, \mathbf{v})$ is a real analytic function on $\mathcal{X}$. $\underline{2)}$ $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q} h_p(\mathbf{x}, \mathbf{x}') < \infty$. $\underline{3)}$ $\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x})\|^2 < \infty$. $\underline{4)}$ $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = 0$.*

Then, for any $J \geq 1$, η-almost surely $\mathrm{FSSD}^2_p(q) = 0$ if and only if $p = q$.

Proof. Since k is real analytic, the components $g_1, \ldots, g_d$ of $\mathbf{g}$ are real analytic by Lemma 15. For each $i = 1, \ldots, d$, if g_i is real analytic, then $\sum_{j=1}^{J} g_i^2(\mathbf{v}_j) = 0$ if and only if $g_i(\mathbf{y}) = 0$ for all $\mathbf{y} \in \mathcal{X}$, η-almost surely (require that the domain $\mathcal{X}$ be a connected open set) [24]. This implies that $\frac{1}{dJ}\sum_{i=1}^{d}\sum_{j=1}^{J} g_i^2(\mathbf{v}_j) = 0$ if and only if $\mathbf{g}(\mathbf{y}) = \mathbf{0}$ for all $\mathbf{y} \in \mathcal{X}$, η-almost surely. By Theorem 14, $\mathbf{g} = \mathbf{0}$ (the zero function) if and only if $p = q$. □

D More on Bahadur Slope

In practice, the main difficulty in determining the approximate Bahadur slope is the computation of $-2\,\mathrm{plim}_{n\to\infty} \frac{\log(1-F(T_n))}{\rho(n)}$, typically requiring the aid of the theory of large deviations. There are further sufficient conditions which make the computation easier. The following conditions are due to [12, 13], first appearing in [1] in a slightly less general form.

Definition 8. Let $\mathcal{D}(a, t)$ be a class of all continuous cumulative distribution functions (CDF) F such that $-2\log(1 - F(x)) = ax^t(1 + o(1))$, as $x \to \infty$ for $a > 0$ and $t > 0$.

Theorem 9 ([12, 13]). *Consider a sequence of test statistic T_n. Assume that*

1. *There exists a function $F(x)$ such that for $\theta \in \Theta_0$, $\lim_{n\to\infty} P_\theta(T_n < x) = F(x)$, for all x, and such that $F \in \mathcal{D}(a, t)$ for some $a > 0$ and $t > 0$ (see Definition 8).*
2. *There exists a continuous, strictly increasing function $R : (0, \infty) \to (0, \infty)$ with $\lim_{n\to\infty} R(n) = \infty$, and a function $b(\theta)$ with $0 < b(\theta) < \infty$ defined on $\Theta\backslash\Theta_0$, such that for all $\theta \in \Theta\backslash\Theta_0$, $\mathrm{plim}_{n\to\infty} T_n/R(n) = b(\theta)$.*

Then, $-2\,\mathrm{plim}_{n\to\infty} \frac{\log(1-F(T_n))}{[R(n)]^t} = a\,[b(\theta)]^t =: c(\theta)$, the approximate slope of the sequence T_n, where $\rho(n) = R(n)^t$ (see Section 4).

Theorem 10 ([12, 13]). *Consider two sequences of test statistics $T_n^{(1)}$ and $T_n^{(2)}$. Let $F^{(i)}$ be the CDF of $T_n^{(i)}$ for $i = 1, 2$. Assume that each sequence satisfies all the conditions in Theorem 9 with $F^{(i)} \in \mathcal{D}(a_i, t_i)$. Further, assume that $\left[R^{(1)}(x)\right]^{t_1} = \left[R^{(2)}(x)\right]^{t_2}$ for all x. Then*

$$\underset{n\to\infty}{\mathrm{plim}} \frac{\log(1 - F^{(1)}(T_n^{(1)}))}{\log(1 - F^{(2)}(T_n^{(2)}))} = \frac{c^{(1)}(\theta)}{c^{(2)}(\theta)} = \varphi_{1,2}(\theta),$$

which is the approximate Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$.

With Theorem 9, the difficulty is in showing that $F \in \mathcal{D}(a, t)$ for some $a > 0, t > 0$. Typically verification of the assumption 2 of Theorem 9 poses no problem. [1] showed that the CDF of $\mathcal{N}(0, 1)$ belongs to $\mathcal{D}(1, 2)$ and the CDF of χ^2_k (chi-squared distribution with k degrees of freedom, fixed k) belongs to $\mathcal{D}(1, 1)$. The following results make it easier to determine whether a given CDF is in the class $\mathcal{D}(a, t)$.

Theorem 11 ([13, Theorem 6, 7]). *Let X have CDF $F \in \mathcal{D}(a, t)$, and $X_1, \ldots, X_m$ be independent random variables, each with CDF $F_i \in \mathcal{D}(a, t)$. Then, the following statements are true.*

1. *If $b > 0$, then the CDF of bX is in $\mathcal{D}(ab^{-t}, t)$.*

2. *$X - b$ has CDF in $\mathcal{D}(a, t)$ provided that $t \geq 1$.*
3. *For $r > 0$, X^r has CDF in $\mathcal{D}(a, r^{-1}t)$ provided that $F(0) = 0$.*
4. *$\max(X_1, \ldots, X_m)$ has CDF in $\mathcal{D}(a, t)$.*
5. *Let $a_1, \ldots, a_m$ be non-negative real numbers such that $a_{max} := \max(a_1, \ldots, a_m) > 0$. Then, $\sum_{i=1}^m a_i X_i$ has CDF in $\mathcal{D}(a \cdot a_{max}^{-t}, t)$ provided that $\sum_{i=1}^m X_i$ has CDF in $\mathcal{D}(a, t)$ and $X_i \geq 0$ for all $i = 1, \ldots, m$.*

E Proof of Theorem 3

Recall Theorem 3:

Theorem 3. *Let $\hat{\boldsymbol{\Sigma}}_q := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}(\mathbf{x}_i)\boldsymbol{\tau}^\top(\mathbf{x}_i) - [\frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}(\mathbf{x}_i)][\frac{1}{n}\sum_{j=1}^n \boldsymbol{\tau}(\mathbf{x}_j)]^\top$ with $\{\mathbf{x}_i\}_{i=1}^n \sim q$. Suppose that the test threshold T_α is set to the $(1-\alpha)$-quantile of the distribution of $\sum_{i=1}^{dJ}(Z_i^2 - 1)\hat{\nu}_i$ where $\{Z_i\}_{i=1}^{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\hat{\nu}_1, \ldots, \hat{\nu}_{dJ}$ are eigenvalues of $\hat{\boldsymbol{\Sigma}}_q$. Then, under H_0, asymptotically the false positive rate is α. Under H_1, for $\{\mathbf{v}_j\}_{j=1}^J$ drawn from a distribution with a density, the test power $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha) \to 1$ as $n \to \infty$.*

Proof. Under H_0, $p = q$ implies that $\hat{\boldsymbol{\Sigma}}_q = \hat{\boldsymbol{\Sigma}}_p$ (empirical estimate of $\boldsymbol{\Sigma}_p$). Let $\lambda_j(A)$ denote the j^{th} eigenvalue of the matrix A. Lemma 16 implies that $A \mapsto \lambda_j(A)$ is continuous on the space of real symmetric matrices, for all j. Since $\mathrm{plim}_{n\to\infty} \|\hat{\boldsymbol{\Sigma}}_p - \boldsymbol{\Sigma}_p\| = 0$, by the continuous mapping theorem, the eigenvalues of $\hat{\boldsymbol{\Sigma}}_p$ converge to the eigenvalues of $\boldsymbol{\Sigma}_p$ in probability. This implies that $\sum_{i=1}^{dJ}(Z_i^2 - 1)\hat{\nu}_i$ converges in probability to $\sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$ as $n \to \infty$, where $\{\omega_i\}_{i=1}^{dJ}$ are eigenvalues of $\boldsymbol{\Sigma}_p$. By Lemma 17, the quantile also converges, and the test threshold thus matches that of the true asymptotic null distribution given in claim 1 of Proposition 2.

Assume H_1 holds. Let $\hat{t}_\alpha, t_\alpha$ be $(1-\alpha)$-quantiles of the distributions of $\sum_{i=1}^{dJ}(Z_i^2 - 1)\hat{\nu}_i$ and $\sum_{i=1}^{dJ}(Z_i^2 - 1)\nu_i$, respectively, where $\{\nu_i\}_{i=1}^{dJ}$ are eigenvalues of $\boldsymbol{\Sigma}_q$. By the same argument as in the previous paragraph, $\hat{t}_\alpha$ converges in probability to t_α, which is a constant independent of the sample size n. Given $\{\mathbf{v}_j\}_{j=1}^J \sim \eta$, where η is a distribution with a density, $\mathrm{FSSD}^2 > 0$ by Theorem 1. It follows that

$$\lim_{n\to\infty} \mathbb{P}\left(n\widehat{\mathrm{FSSD}^2} > \hat{t}_\alpha\right) = \lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathrm{FSSD}^2} - \frac{\hat{t}_\alpha}{n} > 0\right) \overset{(a)}{=} \mathbb{P}\left(\mathrm{FSSD}^2 > 0\right) = 1,$$

where at (a), we use the fact that $\widehat{\mathrm{FSSD}^2}$ converges in probability to FSSD^2 by the law of large numbers, and that $\lim_{n\to\infty} \hat{t}_\alpha/n = 0$. □

F Proof of Theorem 5 (Slope of $n\widehat{\mathrm{FSSD}^2}$)

Recall Theorem 5:

Theorem 5. *The approximate Bahadur slope of $n\widehat{\mathrm{FSSD}^2}$ is $c^{(\mathrm{FSSD})} := \mathrm{FSSD}^2/\omega_1$, where ω_1 is the maximum eigenvalue of $\boldsymbol{\Sigma}_p := \mathbb{E}_{\mathbf{x}\sim p}[\boldsymbol{\tau}(\mathbf{x})\boldsymbol{\tau}^\top(\mathbf{x})]$ and $\rho(n) = n$.*

Proof. We will use Theorem 9 to derive the slope. For the assumption 1 of Theorem 9, we first show that the asymptotic null distribution belongs to the class $\mathcal{D}(a = 1/\omega_1, t = 1)$ as defined in Definition 8. By Proposition 2, the asymptotic null distribution is $\sum_{i=1}^{dJ}\omega_i Z_i^2 - \sum_{i=1}^{dJ}\omega_i$ where $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\omega_1 \geq \cdots \geq \omega_{dJ} \geq 0$ are eigenvalues of $\boldsymbol{\Sigma}_p$. It is known from [1] that the CDF of χ_f^2 is in $\mathcal{D}(1, 1)$ for any fixed degrees of freedom f. Thus, it follows from claim 5 of Theorem 11 that the CDF of $\sum_{i=1}^{dJ}\omega_i Z_i^2$ is in $\mathcal{D}(a = 1/\omega_1, t = 1)$. Claim 2 of Theorem 11 guarantees that the CDF of $\sum_{i=1}^{dJ}\omega_i Z_i^2 - \sum_{i=1}^{dJ}\omega_i$ is in $\mathcal{D}(a = 1/\omega_1, t = 1)$ as desired.

For assumption 2 of Theorem 9, choose $R(n) := n$. It follows from the weak law of large numbers that under H_1, $n\widehat{\mathrm{FSSD}^2}/R(n) \xrightarrow{p} \mathrm{FSSD}^2$. By Theorem 9, the approximate slope is FSSD^2/ω_1. □

G Proof of Theorem 6 (Slope of $\sqrt{n}\widehat{S_l^2}$)

Recall Theorem 6:

Theorem 6. *The approximate Bahadur slope of the linear-time kernel Stein (LKS) test statistic $\sqrt{n}\widehat{S_l^2}$ is $c^{(\mathrm{LKS})} = \frac{1}{2}\frac{\left[\mathbb{E}_q h_p(\mathbf{x},\mathbf{x}')\right]^2}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$, where h_p is the U-statistic kernel of the KSD statistic, and $\rho(n) = n$.*

Proof. We will use Theorem 9 to derive the slope. By the central limit theorem,

$$\sqrt{n}\left(\widehat{S_l^2} - S_p^2(q)\right) \xrightarrow{d} \mathcal{N}(0, 2\mathbb{V}_q[h_p(\mathbf{x},\mathbf{x}')]),$$

where $\mathbb{V}_q[h_p(\mathbf{x},\mathbf{x}')] := \mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}[h_p^2(\mathbf{x},\mathbf{x}')] - (\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}[h_p(\mathbf{x},\mathbf{x}')])^2$. Under $H_0 : p = q$, it follows that $S_p^2(q) = \mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}[h_p(\mathbf{x},\mathbf{x}')] = 0$ by Theorem 14, and $\sqrt{n}\widehat{S_l^2} \xrightarrow{d} \mathcal{N}(0, 2\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')])$ where $\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')] := \mathbb{E}_{\mathbf{x}\sim p}\mathbb{E}_{\mathbf{x}'\sim p}[h_p^2(\mathbf{x},\mathbf{x}')]$. It is known from [1] that the CDF of $\mathcal{N}(0,1)$ is in the class $\mathcal{D}(1,2)$ (see Definition 8). Thus, by property 1 of Theorem 11, the CDF of $\mathcal{N}(0, 2\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')])$ is in $\mathcal{D}\left(a = \frac{1}{2\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')]}, t = 2\right)$.

For assumption 2 of Theorem 9, choose $R(n) := \sqrt{n}$. It follows from the weak law of large numbers that under H_1, $\sqrt{n}\widehat{S_l^2}/R(n) = \widehat{S_l^2} \xrightarrow{p} S_p^2(q)$. By Theorem 9, the approximate slope is $\frac{S_p^4(q)}{2\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')]}$. □

H Proof of Theorem 7

We will first prove a number of useful results that will allow us to prove Theorem 7 at the end. Recall that v denotes a test location in the FSSD test, σ_k^2 denotes the Gaussian kernel bandwidth of the FSSD test, and κ^2 denotes the Gaussian kernel bandwidth of the LKS test.

Proposition 12. *Under the assumption that $J = 1$ (i.e., one test location v), $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$, the approximate Bahadur Slope of $n\widehat{\mathrm{FSSD}^2}$ is*

$$c^{(\mathrm{FSSD})} := \frac{\left(\sigma_k^2\right)^{3/2}\left(\sigma_k^2+2\right)^{5/2} e^{\frac{v^2}{\sigma_k^2+2} - \frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}} \left(\left(\sigma_k^2+1\right)\mu_q + v\left(\sigma_q^2-1\right)\right)^2}{\left(\sigma_k^2+\sigma_q^2\right)^3\left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2+5\right)\sigma_k^2 + 2\right)}. \tag{3}$$

Proof. This result follows directly from Theorem 5 specialized to the case of $p = \mathcal{N}(0,1)$, $q = \mathcal{N}(\mu_q, \sigma_q^2)$, and $J = 1$. Since $dJ = 1$, the covariance matrix

$$\boldsymbol{\Sigma}_p = \mathbb{E}_{x\sim p}\left[\xi_p^2(x,v)\right] = \frac{e^{-\frac{v^2}{\sigma_k^2+2}}\left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2+5\right)\sigma_k^2 + 2\right)}{\sigma_k\left(\sigma_k^2+2\right)^{5/2}}$$

reduces to a scalar, where $\xi_p(x,v) = \left[\frac{\partial}{\partial x}\log p(x)\right]k(x,v) + \frac{\partial}{\partial x}k(x,v) = -e^{-\frac{(v-x)^2}{2\sigma_k^2}}\left(x\sigma_k^2 - v + x\right)/\sigma_k^2$. In this case,

$$\mathrm{FSSD}^2 = \mathbb{E}_{x\sim q}^2\left[\xi_p(x,v)\right] = \frac{\sigma_k^2 e^{-\frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}}\left(\left(\sigma_k^2+1\right)\mu_q + v\left(\sigma_q^2-1\right)\right)^2}{\left(\sigma_k^2+\sigma_q^2\right)^3}.$$

Taking the ratio $\mathrm{FSSD}^2/\mathbb{E}_{x\sim p}\left[\xi_p^2(x,v)\right]$ gives the result. □

Proposition 13. *Assume that $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$. Let $\sqrt{n}\widehat{S_l^2}$ be the linear-time kernel Stein (LKS) test statistic where $\widehat{S_l^2}$ is defined in Section 2 with a Gaussian kernel $k(x,y) = \exp\left(-\frac{(x-y)^2}{2\kappa^2}\right)$. Then, the following statements hold.*

1. *The population kernel Stein discrepancy is*

$$S_p^2(q) = \frac{\mu_q^2\left(\kappa^2+2\sigma_q^2\right)+\left(\sigma_q^2-1\right)^2}{\left(\kappa^2+2\sigma_q^2\right)\sqrt{\frac{2\sigma_q^2}{\kappa^2}+1}}.$$

2. *The approximate Bahadur slope of* $\sqrt{n}\widehat{S_l^2}$ *is*

$$c^{(\mathrm{LKS})} := \frac{\kappa^5\left(\kappa^2+4\right)^{5/2}\left[\mu_q^2\left(\kappa^2+2\sigma_q^2\right)+\left(\sigma_q^2-1\right)^2\right]^2}{2\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)\left(\kappa^2+2\sigma_q^2\right)^3}. \tag{4}$$

3. *Let*

$$c_1^{(\mathrm{LKS})} = \frac{\left(\kappa^2\right)^{5/2}\left(\kappa^2+4\right)^{5/2}\mu_q^4}{2\left(\kappa^2+2\right)\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)}$$

   denote the approximate slope $c^{(\mathrm{LKS})}$ *specialized to when* $q = \mathcal{N}(\mu_q, 1)$*. Then, for any* $\mu_q \neq 0$*, the function* $\kappa^2 \mapsto c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2)$ *is strictly increasing on* $(0, \infty)$*. Further,*

$$\lim_{\kappa^2\to\infty} c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) = \mu_q^4/2. \tag{5}$$

Proof. **Proof of Claim 1, 2**. Recall $\widehat{S_l^2} := \frac{2}{n}\sum_{i=1}^{n/2} h_p(x_{2i-1}, x_{2i})$. With $p = \mathcal{N}(0,1)$, and $k(x,y) = \exp\left(-\frac{(x-y)^2}{2\kappa^2}\right)$, $h_p(x,y)$ can be written as

$$h_p(x,y) := \frac{e^{-\frac{(x-y)^2}{2\kappa^2}}\left(\kappa^2-\left(\kappa^2+1\right)x^2+\left(\kappa^4+2\kappa^2+2\right)xy-\left(\kappa^2+1\right)y^2\right)}{\kappa^4}.$$

By Theorem 6, $c^{(\mathrm{LKS})} = \frac{1}{2}\frac{\left[\mathbb{E}_q h_p(\mathbf{x},\mathbf{x}')\right]^2}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$ which mainly involves expectations with respect to a normal distribution. In computing the expectation $\mathbb{E}_{x'\sim q} h_p(x,x')$, the idea is to form the density for a new normal distribution by combining $\frac{1}{\sqrt{2\pi\sigma_q^2}}e^{-(x-\mu_q)^2/2\sigma_q^2}$ (the density of q) and the term $e^{-\frac{(x-y)^2}{2\kappa^2}}$ in the expression of $h_p(x,y)$. Computation of $\mathbb{E}_{x'\sim q} h_p(x,x')$ will then boil down to computing an expectation wrt. a new normal distribution.

It turns out that

$$\mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}[h_p(x,x')] = \frac{\mu_q^2\left(\kappa^2+2\sigma_q^2\right)+\left(\sigma_q^2-1\right)^2}{\left(\kappa^2+2\sigma_q^2\right)\sqrt{\frac{2\sigma_q^2}{\kappa^2}+1}} = S_p^2(q),$$

$$\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right] = \frac{\left(\kappa^2+4\right)\left(\kappa^4+4\kappa^2+5\right)\kappa^2+12}{\kappa^3\left(\kappa^2+4\right)^{5/2}}.$$

Computing $\frac{1}{2}\frac{S_p^4(q)}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$ gives the slope.

Proof of Claim 3. The expression for $c_1^{(\mathrm{LKS})}$ is obtained straightforwardly by plugging $\sigma_q^2 = 1$ into the expression of $c^{(\mathrm{LKS})}$. Assume $\mu_q \neq 0$. It can be seen that $c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2)$ is differentiable with respect to κ^2 on the interval $(0, \infty)$. The partial derivative is given by

$$\frac{\partial}{\partial\kappa^2}c_1^{(\mathrm{LKS})} = \frac{\left(\kappa^2\right)^{3/2}\left(\kappa^2+4\right)^{3/2}\left(7\kappa^8+56\kappa^6+166\kappa^4+216\kappa^2+120\right)\mu_q^4}{\left(\kappa^2+2\right)^2\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)^2}.$$

Since for any $\mu_q \neq 0$, $\frac{\partial}{\partial\kappa^2}c_1^{(\mathrm{LKS})} > 0$ for $\kappa^2 \in (0, \infty)$, we conclude that $\kappa^2 \mapsto c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2)$ is a strictly increasing function on $(0, \infty)$. By taking the limit, we have $\lim_{\kappa^2\to\infty} c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) = \mu_q^4/2$. □

We are ready to prove Theorem 7. Recall that σ_k^2 is the kernel bandwidth of $n\widehat{\mathrm{FSSD}^2}$, and κ^2 is the kernel bandwidth of $\sqrt{n}\widehat{S_l^2}$ (see Section 2). Recall Theorem 7:

Theorem 7 (Efficiency in the Gaussian mean shift problem)**.** *Let $E_1(\mu_q, v, \sigma_k^2, \kappa^2)$ be the approximate Bahadur efficiency of $n\widehat{\mathrm{FSSD}^2}$ relative to $\sqrt{n}\widehat{S_l^2}$ for the case where $p = \mathcal{N}(0,1), q = \mathcal{N}(\mu_q, 1)$, and $J = 1$ (i.e., one test location v for $n\widehat{\mathrm{FSSD}^2}$). Fix $\sigma_k^2 = 1$ for $n\widehat{\mathrm{FSSD}^2}$. Then, for any $\mu_q \neq 0$, for some $v \in \mathbb{R}$, and for any $\kappa^2 > 0$, we have $E_1(\mu_q, v, \sigma_k^2, \kappa^2) > 2$.*

Proof. By Proposition 12, the approximate slope of $n\widehat{\mathrm{FSSD}^2}$ when $\sigma_q^2 = 1$ is

$$c_1^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2) = \frac{\sigma_k^2 \left(\sigma_k^2 + 2\right)^3 \mu_q^2 e^{\frac{v^2}{\sigma_k^2+2} - \frac{(v-\mu_q)^2}{\sigma_k^2+1}}}{\sqrt{\frac{2}{\sigma_k^2} + 1} \left(\sigma_k^2 + 1\right) \left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2 + 5\right)\sigma_k^2 + 2\right)}.$$

Theorem 10 states that the approximate efficiency $E_1(\mu_q, v, \sigma_k^2, \kappa^2)$ is given by the ratio $\frac{c_1^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2)}{c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2)}$ (see Propositions 12 and 13) of the approximate slopes of the two tests. Pick $\sigma_k^2 = 1$, and for any $\mu_q \neq 0$, pick $v = 2\mu_q$. These choices give the slope

$$c_1^{(\mathrm{FSSD})}(\mu_q, 2\mu_q, 1) = \frac{9\sqrt{3}e^{\frac{5\mu_q^2}{6}}\mu_q^2}{2\left(4\mu_q^2 + 12\right)}.$$

We have

$$\begin{aligned} E_1(\mu_q, v, \sigma_k^2, \kappa^2) &= E_1(\mu_q, 2\mu_q, 1, \kappa^2) \\ &= c_1^{(\mathrm{FSSD})}(\mu_q, 2\mu_q, 1)/c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) \\ &\overset{(a)}{\geq} c_1^{(\mathrm{FSSD})}(\mu_q, 2\mu_q, 1)/\left(\frac{\mu_q^4}{2}\right) \\ &= \frac{9\sqrt{3}e^{\frac{5\mu_q^2}{6}}}{\mu_q^2\left(4\mu_q^2 + 12\right)} := g(\mu_q), \end{aligned}$$

where at (a) we use $c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) \leq \mu_q^4/2$ from (5). It can be seen that for $\mu_q \neq 0$, $g(\mu_q)$ is an even function i.e., $g(\mu_q) = g(-\mu_q)$. The second derivative

$$\frac{\partial^2}{\partial \mu_q^2} g(\mu_q) = \sqrt{3}e^{\frac{5\mu_q^2}{6}} \left(25\mu_q^8 + 45\mu_q^6 - 45\mu_q^4 + 81\mu_q^2 + 486\right) / \left(4\mu_q^4\left(\mu_q^2 + 3\right)^3\right) > 0.$$

To see that $\frac{\partial^2}{\partial \mu_q^2} g(\mu_q) > 0$, consider two cases of $\mu_q^2 \geq 1$ and $0 < \mu_q^2 < 1$. When $\mu_q^2 \geq 1$,

$$g(\mu_q) \geq \sqrt{3}e^{\frac{5\mu_q^2}{6}} \left(25\mu_q^8 + 81\mu_q^2 + 486\right) / \left(4\mu_q^4\left(\mu_q^2 + 3\right)^3\right) > 0,$$

because $45\mu_q^6 - 45\mu_q^4 \geq 0$. When $0 < \mu_q^2 < 1$,

$$g(\mu_q) \geq \sqrt{3}e^{\frac{5\mu_q^2}{6}} \left(25\mu_q^8 + 45\mu_q^6 + 486\right) / \left(4\mu_q^4\left(\mu_q^2 + 3\right)^3\right) > 0,$$

because $-45\mu_q^4 + 81\mu_q^2 \geq 0$. This shows that $g(\mu_q)$ is convex on $(0, \infty)$. The function $g(\mu_q)$ on $\mathbb{R}\backslash\{0\}$ achieves global minima at $\mu_q = \mu_q^* := \pm\sqrt{\frac{3}{10}\left(\sqrt{41} - 1\right)} \approx \pm 1.273$. This implies that

$$\begin{aligned} E_1(\mu_q, v, \sigma_k^2, \kappa^2) &\geq g(\mu_q) \geq g(\mu_q^*) \\ &= \frac{25\sqrt{3}e^{\frac{1}{4}\left(\sqrt{41}-1\right)}}{8\left(\sqrt{41} + 4\right)} \approx 2.00855 > 2. \end{aligned}$$

□

I Known Results

This section presents known results from other works.

Theorem 14 ([9, Theorem 2.2]). *If the kernel k is C_0-universal [6, Definition 4.1], $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}h_p(\mathbf{x},\mathbf{x}') < \infty$, and $\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}} \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\|^2 < \infty$, then $S_p(q) = \|\mathbb{E}_{\mathbf{x}\sim q}\xi_p(\mathbf{x},\cdot)\|_{\mathcal{F}^d} = 0$ if and only if $p = q$.*

Lemma 15 ([8, Lemma 1]). *Let U be an open subset of $\mathbb{R}^d$. If k is a bounded, analytic kernel on $U \times U$, then all functions in the RKHS associated with k are analytic.*[4]

Lemma 16 (Weyl's Perturbation Theorem [4, p. 152]). *Let $\lambda_j(A)$ denote the j^{th} eigenvalue of a square matrix A. If A, B are two Hermitian matrices, then*

$$\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|,$$

where $\|\cdot\|$ denotes the operator norm.

Lemma 17 ([31, Lemma 21.2]). *For any sequence of cumulative distribution functions, $F_n^{-1} \overset{d}{\to} F^{-1}$ if and only if $F_n \overset{d}{\to} F$.*

[4] The result of [8] considers only the case where $U = \mathbb{R}^d$. However, the same proof goes through for any open subset $U \subseteq \mathbb{R}^d$.