Universal Hypothesis Testing with Kernels: Asymptotically Optimal Tests for Goodness of Fit

Shengyu Zhu [1] Biao Chen [2] Pengfei Yang [3] Zhitang Chen [1]

Abstract

We characterize the asymptotic performance of nonparametric goodness of fit testing, otherwise known as the universal hypothesis testing that dates back to Hoeffding (1965). The exponential decay rate of the type-II error probability is used as the asymptotic performance metric, hence an optimal test achieves the maximum decay rate subject to a constant level constraint on the type-I error probability. We show that two classes of Maximum Mean Discrepancy (MMD) based tests attain this optimality on $\mathbb{R}^d$, while a Kernel Stein Discrepancy (KSD) based test achieves a weaker one under this criterion. In the finite sample regime, these tests have similar statistical performance in our experiments, while the KSD based test is more computationally efficient. Key to our approach are Sanov's theorem from large deviation theory and recent results on the weak convergence properties of the MMD and KSD.

1. Introduction

Goodness of fit tests play an important role in machine learning and statistical analysis. Given a model distribution P and sample $\{x_i\}_{i=1}^n := x^n$ originating from an unknown distribution Q, the goal is to decide whether to accept the null hypothesis that Q matches P, or the alternative hypothesis that Q and P are different. Traditional (parametric) approaches may require space partitioning or closed-form integrals (Beirlant et al., 1994; Györfi & Van Der Meulen, 1991; Baringhaus & Henze, 1988; Bowman & Foster, 1993). They become computationally intractable in machine learning applications that involve high dimensional data and complicated models, such as large graphical models and deep generative models (Koller & Friedman, 2009; Salakhutdinov, 2015; Sutherland et al., 2017).

Recently, several efficient tests have been proposed based on Reproducing Kernel Hilbert Space (RKHS) embedding. One is to conduct a Maximum Mean Discrepancy (MMD) based two-sample test by drawing samples from the model distribution P (Lloyd & Ghahramani, 2015). A difficulty with this approach is to determine the number of samples drawn from P relative to n, the sample number of the test sequence. Other tests are based on classes of Stein transformed RHKS functions (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017; Oates et al., 2017; Gorham & Mackey, 2015), where the test statistic is the norm of the smoothness-constrained function with largest expectation under Q, referred to as the Kernel Stein Discrepancy (KSD). The KSD has zero expectation under P and does not require computing integrals or drawing samples. Additionally, constructing explicit features of distributions achieves a linear-time goodness of fit test that is also more interpretable (Jitkrittum et al., 2017).

Despite being efficient and well-behaved in practice, fairly little is known about the statistical optimality of these kernel embedding based tests in a nonparametric setting. Current statistical characterization is limited to consistency, that is, the test type-II error probability decays to zero in the large sample limit under a pre-defined significance level, an upper bound on the type-I error probability. This is established using the convergence of the test statistic to the population statistic (the MMD or KSD between P and Q) under suitable assumptions. Consistency, while being a desired property of statistical tests, does not serve as a meaningful criterion for claiming optimality—a consistent test need not be optimal. An alternative and more insightful approach, one adopted in the current paper, is to study the decay rate of the type-II error probability. However, the current literature lacks such a characterization as the asymptotic distribution of the test statistics either has no closed form (Chwialkowski et al., 2016) or is hard to analyze (Liu et al., 2016; Jitkrittum et al., 2017). In light of the important role and good performance of these kernel tests, the present work seeks an exact characterization on the decay rate of the type-II error probability and further a practically meaningful optimality criterion.

Indeed, there is a long-standing open problem related to op-

[1] Huawei Noah's Ark Lab, Hong Kong, China [2] Syracuse University, Syracuse, NY, USA [3] Cubist Systematic Strategies, New York, NY, USA. Correspondence to: Shengyu Zhu <szhu05@syr.edu>. Part of the work was done when the first and third authors were students at Syracuse University.

timal goodness of fit tests in information theory and statistics (Csiszár & Shields, 2004; Cover & Thomas, 2006), dating back to Hoeffding's test (Hoeffding, 1965). Given a known distribution P, the hypothesis testing $H_0 : x^n \sim P$ and $H_1 : x^n \sim Q$ can be extremely hard when Q is arbitrary but unknown, as opposed to the simple case when Q is known. With independent sample and a known Q, Stein's lemma (cf. Lemma 1) states that the type-II error probability vanishes at most exponentially fast with the exponential decay rate, referred to as the error exponent, being the Kullback-Leibler divergence (KLD) between P and Q. This motivates the so-called universal hypothesis testing problem: *does there exist a nonparametric goodness of fit test that achieves the same optimal error exponent as the simple hypothesis testing problem where Q is known?* Over the years, universally optimal tests only exist when the sample space is finite (Hoeffding, 1965; Unnikrishnan et al., 2011). For continuous sample space like $\mathbb{R}$, attempts have been largely fruitless with the only exception of the works by Zeitouni & Gutman (1991); Yang & Chen (2017). The results were obtained at the cost of a weaker optimality and the proposed test is rather complicated due to the use of Lévy-Prokhorov metric.

In this work, we first show a simple kernel test, comparing the MMD between the reference distribution and the sample empirical distribution with a proper threshold, as an optimal approach to universal hypothesis testing on Polish, locally compact Hausdorff space, e.g., $\mathbb{R}^d$. Taking into account the computation difficulty of non-Gaussian distributions, we further cast the original problem into a two-sample problem as in (Lloyd & Ghahramani, 2015). We show that the same optimality can be attained provided that $\omega(n)$ independent samples are drawn from P. In particular, when distributions are defined on $\mathbb{R}^d$, the biased and unbiased two-sample tests in (Gretton et al., 2012a) with Gaussian kernels can be used to achieve the asymptotic optimality, regardless of the kernel parameters. We then discuss other distance measures for constructing goodness of fit tests. The level constraint on the type-I error probability then becomes difficult to meet for all possible sample sizes. As such, we relax the constraint to an asymptotic one and further show that a KSD based test also achieves the optimal type-II error exponent. Key to our approach is a useful large deviation tool, Sanov's theorem, together with recent results on the weak convergence properties of the MMD (Sriperumbudur, 2016; Simon-Gabriel & Schölkopf, 2016) and the KSD (Gorham & Mackey, 2017).

We remark that the techniques of utilizing Sanov's theorem and the weak convergence property may be of independent interest, and may be used to evaluate other kernel tests.

Other related work. Minimizing the type-II error probability (or equivalently, maximizing test power) subject to a given level has been studied for kernel choice in goodness of fit testing (Jitkrittum et al., 2017) and two-sample testing (Gretton et al., 2012b; Sutherland et al., 2017), based on the asymptotic distribution of the test statistics. The characterization of type-II error probability depends on the sample size as well as the specific kernel for use. Instead, we directly investigate the acceptance region under the alternative hypothesis assisted by Sanov's theorem. Our tests attain the optimal error exponent and are independent of specific kernels as long as they meet the assumptions.

Another criterion, asymptotic Bahadur efficiency, is used for comparing test performance in (Jitkrittum et al., 2017). The proposed linear-time test is shown to always have greater relative efficiency than the linear-time test in (Liu et al., 2016) under a mean-shift alternative, regardless of the choice of parameters for that test. It is not clear whether this claim holds for general alternatives and further if the proposed test is optimal under this criterion.

Paper outline. We begin with a brief review of the MMD and a formal statement of the universal hypothesis testing problem in Section 2. Section 3 presents two classes of kernel tests to universal hypothesis testing and discusses their implications to goodness of fit testing. Section 4 discusses the KSD and other distance measures in constructing goodness of fit tests. All the necessary proofs are provided in Section 5. We perform synthetic experiments in Section 6 to validate our findings and conclude the paper in Section 7.

2. Preliminary and Problem Statement

We briefly review kernel mean embedding and the MMD. We then introduce goodness of fit testing, followed by a formal statement of the universal hypothesis testing.

2.1. Maximum Mean Discrepancy

Let $\mathcal{H}_k$ be a Reproducing Kernel Hilbert Space (RKHS) defined on a topological space $\mathcal{X}$ with reproducing kernel k. Let $\mathcal{P}$ be the set of all Borel probability measures defined on $\mathcal{X}$. The mean embedding of $P \in \mathcal{P}$ in $\mathcal{H}_k$ is a unique element $\mu_k(P) \in \mathcal{H}_k$ such that $\mathbf{E}_{y\sim P} f(y) = \langle f, \mu_k(P)\rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$ (Berlinet & Thomas-Agnan, 2011). We assume that k is bounded continuous, hence the existence of $\mu_k(P)$ is guaranteed by the Riesz representation theorem. The MMD between two probability measures P and Q is defined as the RHKS-distance between their mean embeddings. An expression of the MMD is

$$\begin{aligned}&\mathrm{MMD}[\mathcal{H}_k, P, Q]\\ &= \left(\mathbf{E}_{yy'}k(y,y') + \mathbf{E}_{xx'}k(x,x') - 2\mathbf{E}_{yx}k(y,x)\right)^{1/2},\end{aligned}$$

where y, y' i.i.d. $\sim P$ and x, x' i.i.d. $\sim Q$. In the sequel, we will write $\mathrm{MMD}[\mathcal{H}_k, P, Q] := d_k(P, Q)$. We also refer the reader to a recent overview by Muandet et al. (2017).

If the mean embedding μ_k is an injective map, then the kernel k is said to be characteristic and the MMD d_k becomes a metric on $\mathcal{P}$ (Sriperumbudur et al., 2010). Recently, Simon-Gabriel & Schölkopf (2016, Theorem 55) and Sriperumbudur (2016, Theorem 3.2) have explored a weaker metrizable property of d_k. Throughout the rest of this paper, let $\mathcal{X}$ be a Polish space. Consider the weak topology on $\mathcal{P}$ induced by the weak convergence: $P_l \to P$ weakly if and only if $\mathbf{E}_{y\sim P_l} f(y) \to \mathbf{E}_{y\sim P} f(x)$ for every bounded continuous function $f : \mathcal{X} \to \mathbb{R}$. The following theorem states when d_k metrizes this weak convergence.[1]

Theorem 1 (Simon-Gabriel & Schölkopf (2016); Sriperumbudur (2016))**.** *If $\mathcal{X}$ is Polish, locally compact Hausdorff, and k is continuous and characteristic, then d_k metrizes the weak convergence on $\mathcal{P}$.*

The above theorem indicates that the metric d_k induces the same topology as that of weak convergence. An example of Polish, locally compact Hausdorff space is $\mathbb{R}^d$, and both Gaussian and Laplacian kernels defined on it are continuous and characteristic (Sriperumbudur, 2016).

2.2. Statistical Testing

Given the distribution P and independent sample x^n from an unknown distribution Q, we want to determine whether to accept $H_0 : P = Q$ or $H_1 : P \neq Q$. A decision rule (test) $\Omega(n) = (\Omega_0(n), \Omega_1(n))$ partitions $\mathcal{X}^n$ into two disjoint sets with $\Omega_0(n)\cup\Omega_1(n) = \mathcal{X}^n$. If $x^n \in \Omega_i(n), i = 0, 1$, a decision is made in favor of hypothesis H_i. We say that $\Omega_0(n)$ is an acceptance region for the null hypothesis H_0 and $\Omega_1(n)$ the corresponding critical region. There are two types of errors: a type-I error is made when $P = Q$ is rejected while H_0 is true, and a type-II error occurs when $P = Q$ is accepted despite H_1 being true. The type-I and type-II error probabilities are respectively

$$\alpha_n = P(\Omega_1(n)) = \mathbf{P}_{x^n\sim P}\left(x^n \in \Omega_1(n)\right),$$
$$\beta_n = Q(\Omega_0(n)) = \mathbf{P}_{x^n\sim Q}\left(x^n \in \Omega_0(n)\right).$$

In general, the two types of error probabilities can not be minimized simultaneously. Typical approach is to set an upper bound α on the type-I error probability and consider only level α tests, i.e., tests with $\alpha_n \le \alpha$. A level α test is consistent when it has vanishing type-II error in the large sample limit. Such tests are said to be exponentially consistent if the type-II error probability additionally vanishes exponentially fast with respect to the sample size, i.e., when

$$0 < \liminf_{n\to\infty} -\frac{1}{n}\log\beta_n := \beta < \infty.$$

The above limit is called type-II error exponent. Clearly, $\beta = 0$ implies that the type-II error probability is bounded away from 0 or decays to 0 sub-exponentially, while $\beta = \infty$ indicates it vanishes more than exponentially fast.

2.3. Universal Hypothesis Testing

We first present Stein's lemma on the optimal exponential decay rate of any level α test for simple hypothesis testing between two known distributions. Let $D(P\|Q)$ denote the KLD between two distributions P and Q defined on a Polish space $\mathcal{X}$. The lemma is stated below.

Lemma 1 (Stein's Lemma (Dembo & Zeitouni, 2009; Cover & Thomas, 2006))**.** *Let x^n i.i.d. $\sim R$. Consider hypothesis testing between $H_0 : R = P$ and $H_1 : R = Q$, with $0 < D(P\|Q) < \infty$. Given $0 < \alpha < 1$, let $\Omega^*(n, P, Q) = (\Omega^*_0(n, P, Q), \Omega^*_1(n, P, Q))$ be the optimal level α test with which the type-II error probability is minimized for each n. Then the type-II error probability decays to 0 exponentially at a rate of $D(P\|Q)$ as $n \to \infty$, that is,*

$$\lim_{n\to\infty} -\frac{1}{n}\log Q(\Omega^*_0(n, P, Q)) = D(P\|Q).$$

Let $\Omega(n)$ be a goodness of fit test of level α. When x^n i.i.d. $\sim Q$ under the alternative hypothesis, the corresponding type-II error probability $\beta_n = Q(\Omega_0(n))$ can not be lower than $Q(\Omega^*_0(n, P, Q))$. As a result, Stein's lemma indicates that the type-II error exponent is bounded by $D(P\|Q)$. The problem is to find a test $\Omega(n)$ such that for any given P,

$$\alpha_n \le \alpha,$$
$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n = D(P\|Q),$$

for arbitrary Q satisfying $0 < D(P\|Q) < \infty$, giving rise to the name *universal* hypothesis testing.

3. Maximum Mean Discrepancy Based Tests for Universal Hypothesis Testing

In this section, we study two classes of MMD based goodness of fit tests that are universally asymptotically optimal, followed by discussions on related aspects.

We summarize the assumptions that are made in the last section and that will be used throughout this section: the sample space $\mathcal{X}$ is Polish, locally compact Hausdorff; $\mathcal{P}$ is the set of all Borel probability measures defined on $\mathcal{X}$; kernel k is bounded continuous and characteristic. This ensures that the MMD d_k metrizes weak convergence.

3.1. Simple Kernel Tests

The first test is based on the MMD between the target distribution P and the empirical distribution of sample x^n.

[1] Indeed, Simon-Gabriel & Schölkopf (2016) show that $\mathcal{X}$ only needs to be locally compact Hausdorff. We require $\mathcal{X}$ be Polish in order to utilize some large deviation results.

Though easy to come up with, its optimality for the universal hypothesis testing problem remains unknown.

To proceed, let $\hat{Q}_n$ denote the empirical measure of x^n, i.e., $\hat{Q}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with δ_x being Dirac measure at x. We have a simple kernel test with acceptance region

$$\Omega_0(n) = \left\{ x^n : d_k(P, \hat{Q}_n) \le \gamma_n \right\},$$

where γ_n is a threshold and $d_k^2(P, \hat{Q}_n)$ is computed by

$$\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i, x_j) + \mathbf{E}_{yy'}k(y, y') - \frac{2}{n}\sum_{i=1}^n \mathbf{E}_y k(x_i, y),$$

with y, y' i.i.d. $\sim P$. Throughout this paper, we will simply use $d_k(P, \hat{Q}_n) \le \gamma_n$ to represent the set $\Omega_0(n)$. On the one hand, we want the threshold γ_n to be small so that the test type-II error probability is low; on the other hand, the threshold cannot be too small in order to satisfy the level constraint on the type-I error. The balance between the two error probabilities is attained with a threshold that diminishes at an appropriate rate.

Theorem 2. *For $P \in \mathcal{P}$ and x^n i.i.d. $\sim Q \in \mathcal{P}$, assume that $D(P\|Q) < \infty$. Also assume $0 \le k(\cdot,\cdot) \le K$ with K being a constant value. For a given level α, $0 < \alpha < 1$, set $\gamma_n = \sqrt{K/n}(2 + \sqrt{2\log\alpha^{-1}})$. Then the kernel test $d_k(P, \hat{Q}_n) \le \gamma_n$ is optimal to universal hypothesis testing. That is, when the null hypothesis $H_0 : P = Q$ is true,*

$$P\left(d_k(P, \hat{Q}_n) > \gamma_n\right) \le \alpha;$$

and if the alternative hypothesis $H_1 : P \ne Q$ holds,

$$\liminf_{n\to\infty} -\frac{1}{n} Q\left(d_k(P, \hat{Q}_n) \le \gamma_n\right) = D(P\|Q).$$

A proof is provided in Section 5.1. Similar to (Gretton et al., 2012a), by replacing $\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i, x_j)$ in $d_k^2(P, \hat{Q}_n)$ with $\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i} k(x_i, x_j)$, we can obtain an unbiased test statistic, denoted as $d_u^2(P, \hat{Q}_n)$. We remark that $d_u^2(P, \hat{Q}_n)$ is not a squared quantity and can be negative, due to the unbiasedness. Our result is summarized as follows.

Corollary 1. *Under the same conditions as in Theorem 2, the test $d_u^2(P, \hat{Q}_n) \le \gamma_n^2 + K/n$ is a level α asymptotically optimal test for universal hypothesis testing.*

The tests in this section still require computing integrals, namely, $\mathbf{E}_y k(x_i, y)$ and $\mathbf{E}_{yy'}k(y, y')$. Our purpose here is to show that the universally optimal error exponent is indeed achievable, giving a reasonable criterion to claim optimality for goodness of fit tests as well as a solution to an open problem. In the next section, we consider another class of MMD based tests, achieving this error exponent without the need for computing integrals.

3.2. Kernel Two-Sample Tests

In the context of model criticism, Lloyd & Ghahramani (2015) cast goodness of fit testing into a two-sample problem where one also draws samples y^m from distribution P. A question that arises is the choice of number of samples, which is not obvious due to the lack of an explicit criterion. In light of universal hypothesis testing, we could ask how many samples would suffice to attain the optimal error exponent $D(P\|Q)$.

Denote by $\hat{P}_m$ the empirical measure of y^m. We consider a two-sample test with acceptance region

$$\Omega_0(m, n) = \{(y^m, x^n) : d_k(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}\},$$

where

$$\begin{aligned}\gamma_{m,n} = {} & \left((K/m)^{\frac{1}{2}} + (K/n)^{\frac{1}{2}}\right)\left(2 + \sqrt{2\log(2\alpha^{-1})}\right), \\ d_k^2(\hat{P}_m, \hat{Q}_n) = {} & \sum_{i=1}^m\sum_{j=1}^m k(y_i, y_j) + \sum_{i=1}^n\sum_{j=1}^n k(x_i, x_j) \\ & - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(y_i, x_j),\end{aligned}$$

and K is a finite bound on kernel $k(\cdot,\cdot)$. The statistic $d_k^2(\hat{P}_m, \hat{Q}_n)$ for estimating squared MMD was originally proposed in (Gretton et al., 2012a). Notice that the type-I and type-II error probabilities depend on both P and Q in the two-sample testing. Although additional randomness is introduced, it does not hurt the statistical optimality in terms of the type-II error exponent, as stated below.

Theorem 3. *Assume the same conditions as in Theorem 2, and that y^m i.i.d. $\sim P$ and x^n i.i.d. $\sim Q$. Let $\Omega_1(m, n) = \mathcal{X}^{m+n} \setminus \Omega_0(m, n)$ be the critical region. Then under the null hypothesis $H_0 : P = Q$,*

$$\mathbf{P}_{y^m x^n}(\Omega_1(m, n)) \le \alpha;$$

and under the alternative hypothesis $H_1 : P \ne Q$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbf{P}_{y^m x^n}(\Omega_0(m, n)) = D(P\|Q),$$

provided that $\lim_{n\to\infty}\frac{m}{n} = \infty$.

We present a proof in Section 5.2. Similar to previous simple kernel tests, we can also replace the first two terms in $d_k^2(\hat{P}_m, \hat{Q}_n)$ with $\frac{1}{m(m-1)}\sum_{i=1}^n\sum_{j\ne i} k(y_i, y_j)$ and $\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i} k(x_i, x_j)$. This leads to an unbiased statistic which we denote as $d_u^2(\hat{P}_m, \hat{Q}_n)$ and that has nearly optimal variance (Gretton et al., 2012a). The corollary below shows its universally asymptotic optimality.

Corollary 2. *Under the same assumptions as in Theorem 3, the test $d_u^2(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}^2 + K/m + K/n$ has its type-I error probability below α and type-II error exponent being $D(P\|Q)$, given that $m/n \to \infty$ as $n \to \infty$.*

3.3. Discussions

We make the following remarks on the MMD based tests.

Threshold choice. The distribution-free thresholds used in the MMD based tests are generally too conservative, as the actual distribution P is not taken into account. Alternatively, we may use Monte Carlo or bootstrap methods to empirically estimate the threshold (Gretton et al., 2012a; Chwialkowski et al., 2016; Jitkrittum et al., 2017), making the tests (asymptotically) level α. Notice that these methods introduce additional randomness on the threshold choice and further on the type-II error probability.[2] A simple fix is to take the minimum of the Monte Carlo or bootstrap threshold and the distribution-free threshold, guaranteeing a vanishing threshold and hence the optimal error exponent. In practice, the bootstrap threshold is mostly smaller than the distribution-free threshold.

Finite vs. asymptotic regime. A finitely positive error exponent $\beta = D(P\|Q)$ implies that the error probability decays with $O\left(e^{-n(\beta-\epsilon)}\right)$ where $\epsilon \in (0, \beta)$ can be arbitrarily small. It further implies that kernels in our tests affect only the sub-exponential term in the type-II error probability as long as they are bounded continuous and characteristic. When n is small, the sub-exponential term may dominate and the test performance does depend on kernels. Selecting a proper kernel is an ongoing research topic and we refer the reader to related works such as (Gretton et al., 2012b; Sutherland et al., 2017; Jitkrittum et al., 2017).

Non i.i.d. data. We notice that Chwialkowski et al. (2016) consider non-i.i.d. data by use of wild bootstrap. In general, statistical optimality with non-i.i.d. data, including sample y^m if drawn using the Markov Chain Monte Carlo method, is difficult to establish even for simple hypothesis testing.

Fair alternative. A notion of fair alternative with fixed KLD was raised in (Ramdas et al., 2015) for the two-sample testing as dimension increases. In light of our results or Stein's lemma, the same fair alternative can also be considered for goodness of fit testing.

4. Kernel Stein Discrepancy (KSD) and Other Distance Measures

We discuss other distance measures that may also be used to construct goodness of fit tests in a similar manner to the MMD. We present a sufficient condition for a test to achieve the optimal type-II error exponent under a fixed significance level. The proof follows the same idea of Theorem 2 and is omitted.

Proposition 1. *Consider $\mathcal{X} = \mathbb{R}^d$. Let x^n, P, Q and $\hat{Q}_n$ be assumed in Theorem 2. Let $d(\cdot,\cdot)$ be some metric of weak convergence of probability measures. For a fixed $\alpha \in (0,1)$, a test $\Omega(n) = (\Omega_0(n), \Omega_1(n))$ has level α and is universally asymptotically optimal if*

(a) $P(\Omega_1(n)) \le \alpha$,

(b) $\Omega_0(n) \subset \{x^n : d(P, \hat{Q}_n) \le \gamma_n\}$, $\gamma_n \to 0$ as $n \to \infty$.

For distributions defined on $\mathbb{R}^d$, many other distance measures metrize weak convergence, including Lévy-Prohorov metric, bounded Lipschitz metric, and Wasserstein distance. The total variation distance is an upper bound on the MMD up to a constant (Sriperumbudur et al., 2010), and the KSD can be lower bounded in terms of the MMD or the bounded Lipschitz metric (involving some unknown constants) (Gorham & Mackey, 2017, Theorems 5, 7, and 8). Therefore, comparing these distance measures between P and $\hat{Q}_n$ with a vanishing threshold meets Condition (b).[3] However, to our best knowledge, there does not exist a uniform or distribution-free threshold that makes the tests meet Condition (a). As such, these distance measures do not give rise to asymptotically optimal tests with a fixed level.

We may, nevertheless, relax Condition (a) to an asymptotic case, i.e, $\limsup_{n\to\infty} P(\Omega_1(n)) \le \alpha$. We will focus on the KSD as other distance measures are generally hard to compute in practice. Let p and q be the respective density functions for P and Q defined on $\mathbb{R}^d$. Chwialkowski et al. (2016) and Liu et al. (2016) define the KSD as

$$d_S(P,Q) = \max_{\|f\|_{\mathcal{H}_k}\le 1} \mathbf{E}_{x\sim Q}\left[s_p(x)f(x) + \nabla_x f(x)\right],$$

where $\|f\|_{\mathcal{H}_k} \le 1$ are functions from the unit ball in a RKHS $\mathcal{H}_k$, and $s_p(x) = \nabla_x \log p(x)$ is the score function of $p(x)$. With a C_0-universal kernel (Carmeli et al., 2010) and $\mathbf{E}_{x\sim Q}\|\nabla_x \log p(x) - \nabla_x \log q(x)\|^2 \le \infty$, Chwialkowski et al. (2016, Theorem 2.2) show that $d_S(P,Q) = 0$ if and only if $P = Q$. The squared KSD can also be written as

$$d_S^2(P,Q) = \mathbf{E}_{x\sim Q}\mathbf{E}_{x'\sim Q}\, h_p(x,x'),$$

where $h_p(x,y) = s_p^T(x)s_p(y)k(x,y) + s_p^T(y)\nabla_x k(x,y) + s_p^T(x)\nabla_y k(x,y) + \mathrm{trace}(\nabla_{x,y}k(x,y))$. Then $d_S^2(P,Q)$ can be estimated by $d_S^2(P,\hat{Q}_n) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n h_p(x_i,x_j)$,

[2] We emphasize the subtlety here. That the bootstrap threshold decays to zero in the limit suffices to meet an asymptotic constant constraint, but does not guarantee the desired type-II decay rate. In other words, we also require a decay rate characterization on the bootstrap threshold, which is usually difficult.

[3] Specifically, let d_{BL} and d_S denote the bounded Lipschitz metric and the KSD, respectively. Gorham & Mackey (2017, Theorems 7 and 8) showed that $d_{BL}(P,\hat{Q}_n) \le g(d_S)$ where $g(d_S) \to 0$ as $d_S \to 0$. Thus, there exists γ_n' such that $\{x^n : d_S(P,\hat{Q}_n) \le \gamma_n\} \subset \{x^n : d_{BL}(P,\hat{Q}_n) \le \gamma_n'\}$ and $\gamma_n' \to 0$ as $n \to \infty$.

which is a degenerate V-statistic under the null hypothesis (Chwialkowski et al., 2016).

An acceptance threshold for $d_S^2(P,\hat{Q}_n)$ can be empirically computed through bootstrap (Arcones & Gine, 1992; Chwialkowski et al., 2014; Leucht et al., 2012), and the resulting test satisfies the given level asymptotically. As discussed in Section 3.3, bootstrap introduces randomness on the threshold and further on the type-II error probability, making it difficult to verify Condition (b). Our approach still requires a (deterministically) vanishing threshold that satisfies the asymptotic constraint on the type-I error. Here we use the result of Chwialkowski et al. (2016, Proposition 3.2), which establishes that $nd_S^2(P,\hat{Q}_n)$ converges weakly to some distribution under the null hypothesis. We assume a fixed α-quantile γ_α of the limiting cumulative distribution function, so that $\lim_{n\to\infty} P(d_S^2(P,\hat{Q}_n) > \gamma_\alpha/n) = \alpha$. Thus, if we pick $\gamma_n = o(n)$, e.g., $\gamma_n = (1+\sqrt{\log\alpha^{-1}})n^{-1/2}$, we get $\gamma_n > \gamma_\alpha/n$ in the limit and thus $\lim_{n\to\infty} P(d_S^2(P,\hat{Q}_n) > \gamma_n) \leq \alpha$.

We summarize the above discussions in the following theorem.

Theorem 4. *Let P and Q be defined on $\mathbb{R}^d$, and consider the test $d_S^2(P,\hat{Q}_n) \leq \gamma_n$ with $\gamma_n = (1+\sqrt{\log\alpha^{-1}})n^{-1/2}$.*

(a) If h is Lipschitz continuous, $\mathbf{E}_{x\sim Q} h_p(x,x) < \infty$, and a technical condition on τ-mixing holds (Chwialkowski et al., 2016, Proposition 3.1), then

$$\lim_{n\to\infty} P(d_S^2(P,\hat{Q}_n) > \gamma_n) \leq \alpha.$$

(b) If 1) $d=1$, $k(x,y) = \Phi(x-y)$ for some $\Phi \in C^2$ (twice continuous differentiable) with a non-vanishing generalized Fourier transform; 2) $k(x,y) = \Phi(x-y)$ for some $\Phi \in C^2$ with a non-vanishing generalized Fourier transform, and $(\hat{Q}_n)_{n\geq 1}$ is uniformly tight; 3) $k(x,y) = (c^2 + \|x-y\|_2^2)^\eta$ for $c > 0$ and $\eta \in (-1,0)$ (Gorham & Mackey, 2017, Theorems 5, 7, and 8), then under $H_1 : P \neq Q$ with $D(P\|Q) < \infty$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log Q(d_S^2(P,\hat{Q}_n) \leq \gamma_n) = D(P\|Q).$$

Similar to the distribution-free thresholds in Section 3, γ_n is usually not good enough for finite sample cases. A simple fix is to take the minimum of this threshold and the bootstrap one.

5. Proofs of the Main Results

5.1. Proofs of Simple Kernel Tests

We first present the following lemma for deriving a suitable threshold to make the test satisfy the given level constraint.

Lemma 2. *Assume $0 \leq k(\cdot,\cdot) \leq K$. Given y^m i.i.d. $\sim P$, denote by $\hat{P}_m$ the empirical measure of y^m. It follows that*

$$\mathbf{P}_{y^m}\left(d_k(P,\hat{P}_m) > 2\left(K/m\right)^{\frac{1}{2}} + \epsilon\right) \leq \exp\left(-\frac{\epsilon^2 m}{2K}\right).$$

Proof. Apply Mcdiarmid' inequality and Rademacher average, following the same idea of (Gretton et al., 2012a, Theorem 7). □

We also need two large deviation results: lower semi-continuity of KLD and Sanov's theorem .

Lemma 3 (Van Erven & Harremos (2014)). *For a fixed $Q \in \mathcal{P}$, $D(\cdot\|Q)$ is a lower semi-continuous function with respect to the weak topology of $\mathcal{P}$. That is, for any $\epsilon > 0$, there exists a neighborhood $U \subset \mathcal{P}$ of P such that for any $P' \in U$, $D(P'\|Q) \geq D(P\|Q) - \epsilon$ if $D(P\|Q) < \infty$, and $D(P'\|Q) \to \infty$ as P' tends to P if $D(P\|Q) = \infty$.*

Theorem 5 (Sanov's Theorem (Sanov, 1958; Dembo & Zeitouni, 2009)). *Let x^n i.i.d. $\sim Q \in \mathcal{P}$. For a set $\Gamma \subset \mathcal{P}$,*

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbf{P}_{x^n}(\hat{Q}_n \in \Gamma) \leq \inf_{R\in\operatorname{int}\Gamma} D(R\|Q),$$
$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbf{P}_{x^n}(\hat{Q}_n \in \Gamma) \geq \inf_{R\in\operatorname{cl}\Gamma} D(R\|Q).$$

where $\operatorname{int}\Gamma$ and $\operatorname{cl}\Gamma$ are the interior and closure of Γ with respect to the weak topology on $\mathcal{P}$, respectively.

Outline of proof. Sanov's theorem states that the probability of $\hat{Q}_n \in \Gamma$ vanishes at least exponentially fast if the underlying distribution $Q \notin \operatorname{cl}\Gamma$. Notice that deciding if $x^n \in \{x^n : d_k(P,\hat{Q}_n) \leq \gamma_n\}$ is equivalent to deciding if its empirical measure $\hat{Q}_n \in \{P' : d_k(P,P') \leq \gamma_n\}$. Thus, if Q is eventually excluded by the closure of $\{P' : d_k(P,P') \leq \gamma_n\}$, we get an exponential decay rate of type-II error probability. The rest is to show that the error exponent reaches $D(P\|Q)$ using the weak metrizable property of the MMD and the lower semi-continuity of the KLD.

Proof of Theorem 2. That the test $d_k(P,\hat{Q}_n) \leq \gamma_n$ has level α can be directly seen from Lemma 2. Since Stein's lemma gives an upper bound on the type-II error exponent for any level α test, i.e., $\beta \leq D(P\|Q)$, what remains is to show $\beta \geq D(P\|Q)$.

By our assumption, d_k metrizes weak convergence on $\mathcal{P}$ (cf. Theorem 1). For any constant $\gamma > 0$, there exists an integer n_0 such that $\gamma_n < \gamma$ for all $n > n_0$. Therefore,

$$\begin{aligned}\beta &= \liminf_{n\to\infty} -\frac{1}{n}\log Q\left(d_k(P,\hat{Q}_n) \leq \gamma_n\right)\\ &\geq \liminf_{n\to\infty} -\frac{1}{n}\log Q\left(d_k(P,\hat{Q}_n) \leq \gamma\right)\\ &\geq \inf_{\{P'\in\mathcal{P}: d_k(P,P')\leq\gamma\}} D(P'\|Q), \qquad (1)\end{aligned}$$

where the last inequality is from Sanov's theorem and that $\{P' \in \mathcal{P} : d_k(P, P') \leq \gamma\}$ is closed with respect to weak topology. Since γ can be arbitrary, we have

$$\beta \geq \lim_{\gamma \to 0^+} \inf_{\{P' \in \mathcal{P}: d_k(P,P') \leq \gamma\}} D(P' \| Q).$$

Using the lower semi-continuity of the KLD in Lemma 3 and the assumption that $D(P\|Q) < \infty$, the limit on the right-hand side is greater than $D(P\|Q) - \epsilon$ for arbitrarily given $\epsilon > 0$. This further implies $\beta \geq D(P\|Q)$. □

Proof of Corollary 1. We first have

$$\begin{aligned}
&\left| d_u^2(P, \hat{Q}_n) - d_k^2(P, \hat{Q}_n) \right| \\
=& \left| \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) - \frac{1}{n^2} \sum_{i=1}^{n} k(x_i, x_i) \right| \\
\leq& K/n.
\end{aligned}$$

It holds that

$$\begin{aligned}
\{x^n : d_k^2(P, \hat{Q}_n) \leq \gamma_n^2\} &\subset \{x^n : d_u^2(P, \hat{Q}_n) \leq \gamma_n^2 + K/n\} \\
&\subset \{x^n : d_k^2(P, \hat{Q}_n) \leq \gamma_n^2 + 2K/n\}.
\end{aligned}$$

Thus, under $H_0 : P = Q$, we have

$$\begin{aligned}
P\left(d_u^2(P, \hat{Q}_n) > \gamma_n^2 + K/n\right) &\leq P\left(d_k^2(P, \hat{Q}_n) > \gamma_n^2\right) \\
&\leq \alpha,
\end{aligned}$$

where the last inequality is from Lemma 2 and the fact that $d_k(P, \hat{Q}_n) \geq 0$. The type-II error exponent follows from

$$\begin{aligned}
&\liminf_{n \to \infty} -\frac{1}{n} \log Q\left(d_u^2(P, \hat{Q}_n) \leq \gamma_n^2 + K/n\right) \\
\geq& \liminf_{n \to \infty} -\frac{1}{n} \log Q\left(d_k^2(P, \hat{Q}_n) \leq \gamma_n^2 + 2K/n\right) \\
\geq& D(P\|Q).
\end{aligned}$$

The last inequality can be shown by similar argument of Eq. (1) as $\gamma_n^2 + 2K/n \to 0$ as $n \to \infty$. Applying Stein's lemma completes the proof. □

5.2. Proof of Two-Sample Tests

Proof of Theorem 3. That the two-sample test is level α can be verified by (Gretton et al., 2012a, Theorem 7).

We can write the type-II error probability as

$$\mathbf{P}_{y^m x^n}(\Omega_0(m,n)) = \beta_{m,n}^u + \beta_{m,n}^l,$$

where

$$\begin{aligned}
\beta_{m,n}^u &= \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \leq \gamma_{m,n}, d_k(P, \hat{P}_m) > \gamma'_{m,n}\right), \\
\beta_{m,n}^l &= \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \leq \gamma_{m,n}, d_k(P, \hat{P}_m) \leq \gamma'_{m,n}\right),
\end{aligned}$$

and $\gamma'_{m,n} = 2\sqrt{K/m} + \sqrt{2KnD(P\|Q)/m}$. It suffices to show that $\max\{\beta_{m,n}^u, \beta_{m,n}^l\}$ decreases exponentially as n scales. We first have

$$\beta_{m,n}^u \leq \mathbf{P}_{y^m}\left(d_k(P, \hat{P}_m) > \gamma'_{m,n}\right) \leq e^{-nD(P\|Q)}. \quad (2)$$

The last inequality is also due to Lemma 2. Thus, $\beta_{m,n}^u$ vanishes at least exponentially fast with the error exponent being $D(P\|Q)$.

For $\beta_{m,n}^l$, we have it equal to

$$\begin{aligned}
&\sum_{\{\hat{P}_m : d_k(P, \hat{P}_m) \leq \gamma'_{m,n}\}} P\left(\hat{P}_m\right) Q\left(d_k(\hat{P}_m, \hat{Q}_n) < \gamma_{m,n}\right) \\
\leq& \sup_{\{\hat{P}_m : d_k(P, \hat{P}_m) \leq \gamma'_{m,n}\}} Q\left(d_k(\hat{P}_m, \hat{Q}_n) < \gamma_{m,n}\right) \\
\leq& Q\left(d_k(P, \hat{Q}_n) \leq \gamma_{m,n} + \gamma'_{m,n}\right),
\end{aligned}$$

where the last inequality is because d_k is a metric. Similar to Eq. (1), we get

$$\liminf_{n \to \infty} -\frac{1}{n} \log \beta_{n,m}^l \geq D(P\|Q),$$

because $\gamma_{m,n} + \gamma'_{m,n} \to 0$ as $n \to \infty$. Together with Eq. (2), we have under $H_1 : P \neq Q$,

$$\liminf_{n \to \infty} -\frac{1}{n} \log \mathbf{P}_{y^m x^n}\left(\Omega_0(m,n)\right) \geq D(P\|Q).$$

We next show the other direction under H_1. We have

$$\begin{aligned}
&\mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_m) \leq \gamma_{m,n}\right) \\
\overset{(a)}{\geq}& \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, P) \leq \gamma'_m, d_k(P, \hat{Q}_n) \leq \gamma'_n\right) \\
=& P\left(d_k(\hat{P}_m, P) \leq \gamma'_m\right) Q\left(d_k(P, \hat{Q}_n) \leq \gamma'_n\right),
\end{aligned}$$

where (a) is because d_k is a metric, and we choose $\gamma'_m = (2 + \sqrt{2\log(2\alpha^{-1})})\sqrt{K/m}$ and $\gamma'_n = (2 + \sqrt{2\log(2\alpha^{-1})})\sqrt{K/n}$ so that $\gamma_{m,n} = \gamma'_m + \gamma'_n$. Then Lemma 2 implies $P(d_k(\hat{P}_m, P) \leq \gamma'_m) > 1 - \alpha/2$ and $P(d_k(P, \hat{Q}_n) \leq \gamma'_n) > 1 - \alpha/2$. Together with Stein's Lemma, we get

$$\begin{aligned}
&\liminf_{n \to \infty} -\frac{1}{n} \log \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \leq \gamma_{m,n}\right) \\
\leq& \liminf_{n \to \infty} -\frac{1}{n} \log\left((1 - \alpha/2)\, Q\left(d_k(P, \hat{Q}_n) \leq \gamma'_n\right)\right) \\
\leq& D(P\|Q).
\end{aligned}$$

□

Proof of Corollary 2. We have

$$\left| d_u^2(\hat{P}_m, \hat{Q}_n) - d_k^2(\hat{P}_m, \hat{Q}_n) \right| \leq K/n + K/m.$$

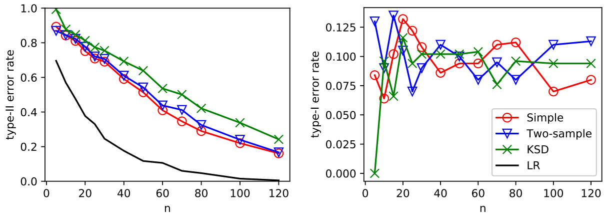
The rest follows the same idea of Corollary 1, along with (Gretton et al., 2012a, Theorem 7) □

6. Experiments

We present empirical results to validate our theoretic findings. We note that there have been extensive experiments on performance comparisons of the MMD based two-sample tests and the KSD based tests in (Chwialkowski et al., 2016; Liu et al., 2016; Jitkrittum et al., 2017), where the number m of samples drawn from P is usually fixed. Our focus will be to compare statistical performance of KSD based tests and the tests of $d_k(P, \hat{Q}_n)$ and $d_k(\hat{P}_m, \hat{Q}_n)$ in the finite sample regime.

We evaluate the following tests with a fixed level $\alpha = 0.1$, all using Gaussian kernel $k(x, y) = e^{-\|x-y\|_2^2/(2w)}$: 1) `Simple`: the simple kernel test $d_k(P, \hat{Q}_n)$. The threshold is estimated by drawing i.i.d. samples under the model P, i.e., the Monte Carlo method. The number of trials is 500. 2) `Two-sample`: the two-sample test $d_k(\hat{P}_m, \hat{Q}_n)$ with $m = n^{1.5}$. Threshold is obtained from the bootstrap method in (Gretton et al., 2012a), with 500 bootstrap replicates. 3) `KSD`: a KSD based test. We simply use the test proposed in (Chwialkowski et al., 2016), as other KSD based tests (Liu et al., 2016; Jitkrittum et al., 2017) have comparable performance in terms of the type-II error probability. We use wild bootstrap method (Chwialkowski et al., 2016) with 500 replicates to estimate the α-quantile.

Gaussian vs. Laplace. We consider a one-dimensional problem in which $P : \mathcal{N}(0, 2\sqrt{2})$ and Q : Laplace$(0, 2)$, a zero-mean Laplace distribution with scale parameter 2. The parameters are chosen so that P and Q have the same mean and variance. We repeat 500 trials of each hypothesis with respect to different sample sizes, and pick a fixed bandwidth $w = 1$ for all the kernel tests. We also evaluate the likelihood ratio test `LR`, an oracle approach assuming both P and Q are known. In Figure 1a, the test `LR` has the lowest type-II error rate as expected, while `Simple` and `Two-sample` perform slightly better than `KSD`. All the kernel tests have the type-I error rates around the given level $\alpha = 0.1$, shown in Figure 1b.



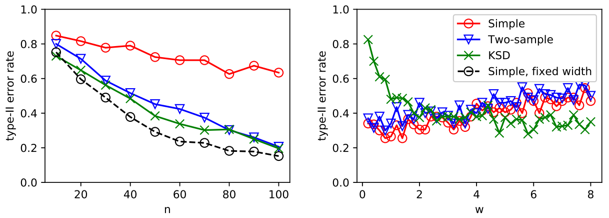
(a)

(b)

Figure 1. Gaussian vs. Laplace.

Gaussian Mixture. We next consider a similar experiment setting to (Liu et al., 2016). We draw i.i.d. sample x^n from $Q : \sum_{k=1}^{5} a_k \mathcal{N}(x; \mu_k, \sigma^2)$ with $a_k = 1/5, \sigma^2 = 1$, and μ_k randomly drawn from Uniform$[0, 10]$. We then generate P by adding standard Gaussian noise (perturbation) to μ_k. In (Liu et al., 2016), the number of samples y^m drawn from P is fixed while varying the observed sample number n. Here we pick $m = n^{1.5}$ and report the type-II error rates in Figure 2, averaged over 500 random trials.

With the median heuristic for bandwidth choice, `KSD` and `Two-sample` perform similarly whereas `Simple` has its type-II error probability decreasing slowly, as shown in Figure 2a. Picking a fixed bandwidth $w = 1$ for `Simple` again achieves a better performance. In light of the role of kernels, we then search over the kernel bandwidths in $[0, 8]$ for a fixed sample size $n = 50$. In Figure 2b, `Simple` and `Two-smaple` tend to have lower type-II error rates when w is small, while `KSD` achieves lower error rates around $w = 5$. The optimal type-II error rates of `Simple` and `KSD` are close and are slightly lower than that of `Two-sample`. Drawing more samples y^m from P may reduce the gap but also increases the computation cost. Besides, we observe that `KSD` is more computationally efficient in this experiment, as it does not need to draw samples.



(a)

(b)

Figure 2. Gaussian mixture. (a) median bandwidth for `Simple`, `Two-sample`, and `KSD`, and a fixed bandwidth $w = 1$ for `Simple`; (b) fixing $n = 50$ and varying kernel bandwidths.

7. Concluding Remarks

In this paper, we have shown that two classes of MMD based tests are universally asymptotically optimal for goodness of fit testing, and that a KSD based test achieves a weaker optimality in the sense that a relaxed level constraint is placed on the type-I error probability. In the finite sample regime, these kernel tests have similar performance, while the KSD based test is more efficient. Our work not only solves a long-standing open problem in information theory and statistics, but also provides theoretic guarantee for these kernel tests in terms of statistical performance. We believe that the technique of using Sanov's theorem and the weak convergence properties of the MMD and the KSD can be further used to evaluate other kernel tests.

References

Arcones, Miguel A and Gine, Evarist. On the bootstrap of U and V statistics. *The Annals of Statistics*, pp. 655–674, 1992.

Baringhaus, L and Henze, N. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

Beirlant, Jan, Györfi, László, and Lugosi, Gábor. On the asymptotic normality of the L_1- and L_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22(3):309–318, 1994.

Berlinet, Alain and Thomas-Agnan, Christine. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

Bowman, AW and Foster, PJ. Adaptive smoothing and density-based tests of multivariate normality. *Journal of the American Statistical Association*, 88(422):529–537, 1993.

Carmeli, Claudio, De Vito, Ernesto, Toigo, Alessandro, and Umanitá, Veronica. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Chwialkowski, Kacper, Strathmann, Heiko, and Gretton, Arthur. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2016.

Chwialkowski, Kacper P, Sejdinovic, Dino, and Gretton, Arthur. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, 2014.

Cover, Thomas M and Thomas, Joy A. *Elements of Information Theory*. New York: Wiley, 2nd edition, 2006.

Csiszár, Imre and Shields, Paul C. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

Dembo, Amir and Zeitouni, Ofer. *Large Deviations Techniques and Applications*. New York: Springer, 2009.

Gorham, Jackson and Mackey, Lester. Measuring sample quality with Stein's method. In *NIPS*, 2015.

Gorham, Jackson and Mackey, Lester. Measuring sample quality with kernels. In *ICML*, 2017.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012a.

Gretton, Arthur, Sejdinovic, Dino, Strathmann, Heiko, Balakrishnan, Sivaraman, Pontil, Massimiliano, Fukumizu, Kenji, and Sriperumbudur, Bharath K. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, 2012b.

Györfi, László and Van Der Meulen, Edward C. A consistent goodness of fit test based on the total variation distance. In *Nonparametric Functional Estimation and Related Topics*, pp. 631–645. Springer, 1991.

Hoeffding, Wassily. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pp. 369–401, 1965.

Jitkrittum, Wittawat, Xu, Wenkai, Szabo, Zoltan, Fukumizu, Kenji, and Gretton, Arthur. A linear-time kernel goodness-of-fit test. In *NIPS*, 2017.

Koller, Daphne and Friedman, Nir. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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

Liu, Qiang, Lee, Jason, and Jordan, Michael. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, 2016.

Lloyd, James Robert and Ghahramani, Zoubin. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

Muandet, Krikamol, Fukumizu, Kenji, Sriperumbudur, Bharath, and Schlkopf, Bernhard. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Oates, Chris J, Girolami, Mark, and Chopin, Nicolas. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

Ramdas, Aaditya, Reddi, Sashank Jakkam, Póczos, Barnabás, Singh, Aarti, and Wasserman, Larry A. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, 2015.

Salakhutdinov, Ruslan. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015.

Sanov, Ivan N. On the probability of large deviations of random variables. Technical report, North Carolina State University. Dept. of Statistics, 1958.

Simon-Gabriel, C.-J. and Schölkopf, B. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *arXiv preprint arXiv:1604.05251*, 2016.

Sriperumbudur, Bharath. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 08 2016.

Sriperumbudur, Bharath K, Gretton, Arthur, Fukumizu, Kenji, Schölkopf, Bernhard, and Lanckriet, Gert RG. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11 (Apr):1517–1561, 2010.

Sutherland, D., Tung, H-Y, Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017.

Unnikrishnan, Jayakrishnan, Huang, Dayu, Meyn, Sean P, Surana, Amit, and Veeravalli, Venugopal V. Universal and composite hypothesis testing via mismatched divergence. *IEEE Transactions on Information Theory*, 57(3): 1587–1603, 2011.

Van Erven, Tim and Harremos, Peter. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Yang, Pengfei and Chen, Biao. Robust Kullback-Leibler divergence and universal hypothesis testing for continuous distributions. *arxiv preprint arxiv:1711.04238*, 2017.

Zeitouni, Ofer and Gutman, Michael. On universal hypotheses testing via large deviations. *IEEE Trans. Inf. Theory*, 37(2):285–290, 1991.