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

Shengyu Zhu
Huawei Noah's Ark Lab
Hong Kong

Biao Chen
Syracuse University
Syracuse, NY

Pengfei Yang
Cubist Systematic Strategies
New York, NY

Zhitang Chen
Huawei Noah's Ark Lab
Hong Kong

Abstract

We characterize the asymptotic performance of nonparametric goodness of fit testing. The exponential decay rate of the type-II error probability is used as the asymptotic performance metric, and a test is optimal if it achieves the maximum 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 the quadratic-time Kernel Stein Discrepancy (KSD) based tests achieve the maximum exponential decay rate under a relaxed level constraint. Under the same performance metric, we proceed to show that the quadratic-time MMD based two-sample tests are also optimal for general two-sample problems, provided that kernels are bounded continuous and characteristic. Key to our approach are Sanov's theorem from large deviation theory and the weak metrizable 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 $x^n := \{x_i\}_{i=1}^n$ 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 [6, 7, 9, 27]. They become computationally intractable to machine learning applications that involve high dimensional data and complicated models [30, 39, 46].

Recently, several efficient tests have been proposed based on Reproducing Kernel Hilbert Space (RKHS) embedding [36, 43]. One is to conduct a Maximum Mean Discrepancy (MMD) based two-sample test by drawing samples from the model distribution P [35]. 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 RKHS functions [12, 22, 23, 34, 37], where the test statistic is the norm of the smoothness-constrained function with the largest expectation under Q and is referred to as the Kernel Stein Discrepancy (KSD). The KSD based tests only require knowing the density function of P up to the normalization constant, and do not need to compute integrals or draw samples. Additionally, constructing explicit features of distributions results in a linear-time goodness-of-fit test that is also more interpretable [29].

Motivated by their good performance in practice, this paper investigates the statistical optimality of these kernel based goodness-of-fit tests, a long-standing open problem in information theory and statistics [15, 17, 28]. Given distribution P, the hypothesis testing between $H_0 : x^n \sim P$ 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, the type-II error probability of an optimal test vanishes exponentially fast w.r.t. the sample size n, and the exponential decay rate coincides with the Kullback-Leibler Divergence (KLD) between P and Q (cf. Lemma 1). This motivates the so-called universal hypothesis testing problem, originally proposed by Hoeffding [28]: *does there exist a nonparametric goodness-of-fit test that achieves the same optimal exponential decay rate as in the simple hypothesis testing problem where Q is known?* Over the years, universally optimal tests only exist when the sample space is finite, i.e., when P and Q are both multinomial [28, 49]. For a more general sample space, attempts have been largely fruitless with the only exception of [53, 51].

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Their results, however, were obtained at the cost of a weaker optimality and the proposed tests are rather complicated due to use of Lévy-Prokhorov metric. We remark that even the existence of such a test remains unknown when the sample space is non-finite.

Contributions. We first show a simple kernel test, comparing the MMD between the target distribution and the sample empirical distribution with a proper threshold, as an optimal approach to the universal hypothesis testing problem when the sample space is Polish, locally compact Hausdorff, e.g., $\mathbb{R}^d$. To the best of our knowledge, this is the first result on the universal optimality for a general, non-finite sample space. Taking into account the difficulty of obtaining closed-form integrals for non-Gaussian distributions, we then follow [35] to cast the original problem into a two-sample problem. We establish the same optimality for the quadratic-time kernel two-sample tests proposed in [25], provided that $\omega(n)$ independent samples are drawn from P. For the KSD based tests, the constant level constraint on the type-I error probability is difficult to satisfy for all possible sample sizes. By relaxing the constraint to an asymptotic one and assuming additional conditions, we establish the optimal exponential decay rate of the type-II error probability for the quadratic-time KSD based tests proposed in [12, 34].

As another contribution, we proceed to investigate the quadratic-time kernel two-sample tests in a more general setting where the sample sizes scale in the same order, e.g., when the two sets of samples have the same size. We show that the type-II error probability also vanishes exponentially fast. The obtained exponential decay rate is further shown to be optimal among all two-sample tests under the same level constraint, and is independent of particular kernels provided that they are bounded continuous and characteristic.

Key to our approach are Sanov's theorem from large deviation theory [19] and the weak metrizable properties of the MMD [42, 44] and the KSD [23], which enable us to directly investigate the acceptance region defined by the test, rather than using the test statistic as an intermediate.

Paper Outline. Section 2 introduces the asymptotic statistical criterion used in this paper and formally states the problem of universal hypothesis testing. Section 3 reviews related works. In Section 4, we present two classes of MMD based tests that are optimal for universal hypothesis testing and discuss their implications to goodness of fit testing. Section 5 considers the KSD based goodness-of-fit tests and Section 6 establishes the universal optimality of the quadratic-time MMD based two-sample tests in a more general setting. We conclude this paper in Section 7.

2 Problem

Throughout this paper, let $\mathcal{X}$ be a Polish space (i.e., a separable completely metrizable topological space) and $\mathcal{P}$ the set of Borel probability measures defined on $\mathcal{X}$. Given a distribution $P \in \mathcal{P}$ and sample x^n from an unknown distribution $Q \in \mathcal{P}$, we want to determine whether to accept $H_0 : P = Q$ or $H_1 : P \neq Q$. A 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 rejection region. 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 two error probabilities are $P(\Omega_1(n)) := \mathbf{P}_{x^n \sim P}(x^n \in \Omega_1(n))$ and $Q(\Omega_0(n)) := \mathbf{P}_{x^n \sim Q}(x^n \in \Omega_0(n))$ with $Q \neq P$, respectively.

In general, the two error probabilities can not be minimized simultaneously. A commonly used approach, the so-called Neyman-Pearson approach [11], is to set an upper bound α on the type-I error probability and considers only level α tests, i.e., tests with $P(\Omega_1(n)) \leq \alpha$. However, similar to the two-sample problem [25], it is not possible to distinguish distributions with high probability at a given, fixed sample, without prior assumptions on the difference between P and Q. We therefore consider an asymptotic statistical criterion as the performance metric.

A level α test is said to be consistent if the type-II error probability vanishes in the large sample limit. Such a test is exponentially consistent when the error probability additionally vanishes exponentially fast w.r.t. the sample size, that is, when

$$\liminf_{n \to \infty} -\frac{1}{n} \log Q(\Omega_0(n)) > 0.$$

The above limit is also referred to as the type-II error exponent in information theory. Clearly, the larger the error exponent, the faster the error probability decreases in the sample limit. Under this criterion, an optimal test would achieve the maximum type-II error exponent while satisfying the level constraint. Error exponent is a widely used metric in source coding and channel coding [15], and is closely related to two other asymptotic statistical criteria [41]. In particular, the Chernoff index equals the minimum of the type-I and type-II error exponents, and the exact Bahadur slope is equivalent to twice of the type-I error exponent with a constant constraint on the type-II error probability.

We present a useful lemma which gives the optimal type-II error exponent of any level α test for simple

hypothesis testing between two known distributions. Let $D(P\|Q)$ denote the KLD between P and Q. That is, $D(P\|Q) = \mathbf{E}_P \log(dP/dQ)$ where dP/dQ stands for the Radon-Nikodym derivative of P w.r.t. Q when it exists, and $D(P\|Q) = \infty$ otherwise [19].

Lemma 1 (Chernoff-Stein Lemma [15, 19])**.** *Let x^n i.i.d. $\sim R$. Consider simple hypothesis testing between $H_0 : R = P \in \mathcal{P}$ and $H_1 : R = Q \in \mathcal{P}$, 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. It follows that*

$$\lim_{n\to\infty} -\frac{1}{n} \log Q(\Omega_0^*(n, P, Q)) = D(P\|Q).$$

Problem Statement. Let $\Omega(n) = \{\Omega_0(n), \Omega_1(n)\}$ be a nonparametric goodness-of-fit test of level α. With x^n i.i.d. $\sim Q$ under the alternative hypothesis, the corresponding type-II error probability $Q(\Omega_0(n))$ can not be lower than $Q(\Omega_0^*(n, P, Q))$. As such, Chernoff-Stein lemma indicates that its type-II error exponent is bounded by $D(P\|Q)$. For any given P, the problem is to find a goodness-of-fit test $\Omega(n)$, if it exists, so that

1. *under $H_0 : P = Q$, $\mathbf{P}_{x^n}(\Omega_1(n)) \le \alpha$,*
2. *under $H_1 : P \neq Q$,*

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

for arbitrary Q with $0 < D(P\|Q) < \infty$,

giving rise to the name *universal* hypothesis testing.

3 Related Work

The decay rate of the type-II error probability has been widely investigated for existing kernel based tests. For the simple kernel tests in [1, 47, 48] and the kernel two-sample tests in [14, 21, 24, 26, 46, 52], analysis is based on the test statistics, through their asymptotic distributions or some probabilistic bounds on their convergence to the population statistics. The resulting characterizations depend on kernels and are loose in general. For the KSD based tests, current statistical characterization is limited to consistency; the asymptotic distributions of the test statistics either have no closed form [12] or are hard to analyze [29, 34].

Other asymptotic statistical criteria have also been used for comparing nonparametric goodness-of-fit tests. Jitkrittum et al. [29] used the approximate Bahadur slope and showed that their linear-time test has greater relative efficiency than the linear-time test proposed in [34], assuming a mean-shift alternative. However, it is not clear whether such a result holds for a more general alternative. Balasubramanian et al. [5] investigated the detection boundary and showed that the simple kernel test is suboptimal under this criterion. A minimax optimal test was then proposed for a composite alternative, where the worst-case performance w.r.t. a set of probability measures is optimized. In contrast, our optimality criterion is much stronger in that the optimality must hold for any distribution defining the alternative hypothesis; specifically, the nonparametric test must achieve the maximum type-II error exponent $D(P\|Q)$ for any Q satisfying $0 < D(P\|Q) < \infty$.

4 Maximum Mean Discrepancy Based Goodness-of-Fit Tests

This section studies two classes of MMD based tests for universal hypothesis testing, followed by discussions on related aspects. We begin with a brief review of the MMD and of Sanov's theorem.

Let $\mathcal{H}_k$ be an RKHS defined on $\mathcal{X}$ with reproducing kernel k. 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 $f \in \mathcal{H}_k$ [8]. 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 RKHS-distance between their mean embeddings, which can be expressed as

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

where y, y' i.i.d. $\sim P$ and x, x' i.i.d. $\sim Q$.

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 $\mathcal{P}$ [45]. A weak metrizable property of d_k has also been established recently. Consider the weak topology on $\mathcal{P}$ induced by the weak convergence: a sequence of probability measures $P_l \to P$ weakly if and only if $\mathbf{E}_{y\sim P_l} f(y) \to \mathbf{E}_{y\sim P} f(y)$ 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 ([42, Theorem 55], [44, Theorem 3.2])**.** *If $\mathcal{X}$ is Polish, locally compact Hausdorff, and k is continuous and characteristic, then d_k metrizes the weak convergence on $\mathcal{P}$.*

We note that the weak metrizable property is also favored for training deep generative models [3, 4, 32].

[1] Indeed, Simon-Gabriel and Schölkopf [42] show that $\mathcal{X}$ only needs to be locally compact Hausdorff. We require $\mathcal{X}$ be Polish in order to utilize Sanov's theorem.

An example of Polish, locally compact Hausdorff space is $\mathbb{R}^d$, and both Gaussian and Laplacian kernels defined on it are bounded continuous and characteristic [44].

We next introduce Sanov's theorem from large deviation theory, which, together with the weak metrizable property of the MMD, is critical to establish our main results in this section. Denote by $\hat{Q}_n$ 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 the Dirac measure at x.

Theorem 2 (Sanov's Theorem [40, 19])**.** *Let x^n i.i.d. $\sim Q \in \mathcal{P}$. For a set $\Gamma \subset \mathcal{P}$, it holds that*

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

where $\operatorname{int}\Gamma$ *and* $\operatorname{cl}\Gamma$ *are the interior and closure of* Γ *w.r.t. the weak topology on* $\mathcal{P}$*, respectively.*

Sanov's theorem states that if the underlying distribution Q is not in $\operatorname{cl}\Gamma$, the closure of a set Γ of distributions, then the probability of its empirical distribution $\hat{Q}_n$ lying in $\operatorname{cl}\Gamma$ goes to 0 at least exponentially fast. This enables us to directly investigate type-II error exponent through the empirical distribution and the acceptance region, rather than through the limiting performance of the test statistics. Moreover, the lower bound on the error exponent would establish the universal optimality if it is no lower than $D(P\|Q)$ for a goodness-of-fit test.

We now state the two classes of MMD based goodness-of-fit tests that are universally optimal.

4.1 Simple Kernel Tests

The first test directly computes the MMD between the target distribution P and the empirical distribution of sample x^n. Though having been investigated in [1, 5, 47, 48], its optimality for the universal hypothesis testing problem remains unknown.

Let also $\hat{Q}_n$ be the empirical measure of x^n. 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 represents a threshold and $d_k^2(P, \hat{Q}_n)$ equals

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

with y, y' i.i.d. $\sim P$. 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 meet the level constraint on the type-I error probability. The balance between the two error probabilities is attained with a threshold that diminishes at an appropriate rate.

Theorem 3. *Let $\mathcal{X}$ be Polish, locally compact Hausdorff. For $P \in \mathcal{P}$ and x^n i.i.d. $\sim Q \in \mathcal{P}$, assume $0 < D(P\|Q) < \infty$ under the alternative hypothesis H_1. Further assume that kernel k is bounded continuous and characteristic, with $0 \le k(\cdot,\cdot) \le K$ for some $K > 0$. For a given α, $0 < \alpha < 1$, set $\gamma_n = \sqrt{2K/n}\left(1 + \sqrt{-\log\alpha}\right)$. Then the simple kernel test $d_k(P, \hat{Q}_n) \le \gamma_n$ is an optimal level α test for the universal hypothesis testing problem, that is,*

1. *under $H_0 : P = Q$, $\mathbf{P}_{x^n}\left(d_k(P, \hat{Q}_n) > \gamma_n\right) \le \alpha$,*
2. *under $H_1 : P \ne Q$,*

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

Proof. That $d_k(P, \hat{Q}_n) \le \gamma_n$ has level α can be directly verified by [48, Eq. (24)] (see Lemma 2 in Appendix A). Let $\beta = \liminf_{n\to\infty} -\frac{1}{n}\log \mathbf{P}_{x^n}(d_k(P, \hat{Q}_n) \le \gamma_n)$ under H_1. According to Chernoff-Stein lemma, we only need to show $\beta \ge D(P\|Q)$.

To apply Sanov's theorem, we notice that deciding if $x^n \in \{x^n : d_k(P, \hat{Q}_n) \le \gamma_n\}$ is equivalent to deciding if its empirical measure $\hat{Q}_n \in \{P' : d_k(P, P') \le \gamma_n\}$. Since $\gamma_n \to 0$ as $n \to \infty$, for any constant $\gamma > 0$, there exists an integer n_0 such that $\gamma_n < \gamma$ for all $n > n_0$. Hence, $\{P' : d_k(P, P') \le \gamma_n\} \subset \{P' : d_k(P, P') \le \gamma\}$ for large enough n. It follows that for any $\gamma > 0$,

$$\begin{aligned}\beta &\ge \liminf_{n\to\infty} -\frac{1}{n}\log \mathbf{P}_{x^n}\left(d_k(P, \hat{Q}_n) \le \gamma\right) \\ &\ge \inf_{\{P'\in\mathcal{P}: d_k(P,P')\le\gamma\}} D(P'\|Q), \end{aligned} \tag{1}$$

where the last inequality is from Sanov's theorem and that $\{P' \in \mathcal{P} : d_k(P, P') \le \gamma\}$ is closed w.r.t. the weak topology (cf. Theorem 1). Then for any given $\epsilon > 0$, there exists some $\gamma > 0$ such that $\inf_{\{P'\in\mathcal{P}: d_k(P,P')\le\gamma\}} D(P'\|Q) \ge D(P\|Q) - \epsilon$, using the lower semi-continuity of the KLD [50] (Lemma 3 in Appendix A) and the assumption that $0 < D(P\|Q) < \infty$ under H_1. This further implies $\beta \ge D(P\|Q)$. □

It is worth noting that we simply select one threshold γ_n in the above theorem. Indeed, any vanishing threshold $\gamma_n' > 0$ with $\gamma_n' \ge \gamma_n$ leads to the same optimality w.r.t. the type-II error exponent, an asymptotic statistical criterion. A larger threshold, however, may result in a higher type-II error probability in the finite sample regime. A further discussion on the threshold choice will be given in Section 4.3.

The test statistic $d_k^2(P, \hat{Q}_n)$ is a biased estimator of $d_k^2(P, Q)$. By replacing $\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i, x_j)$ with $\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i} k(x_i, x_j)$, we obtain an unbiased statistic denoted as $d_u^2(P, \hat{Q}_n)$. We comment that $d_u^2(P, \hat{Q}_n)$ is not a squared quantity and can be negative, due to the unbiasedness. The following result shows that $d_u^2(P, \hat{Q})$ can also be used to construct a universally optimal test.

Corollary 1. *Under the same conditions of Theorem 3, the test $d_u^2(P, \hat{Q}_n) \leq \gamma_n^2 + K/n$ is a level α asymptotically optimal test for universal hypothesis testing.*

Proof (sketch). As $0 \leq k(\cdot,\cdot) \leq K$, we get $\{x^n : d_k^2(P, \hat{Q}_n) \leq \gamma_n^2\} \subset \{x^n : d_u^2(P, \hat{Q}_n) \leq \gamma_n^2 + K/n\} \subset \{x^n : d_k^2(P, \hat{Q}_n) \leq \gamma_n^2 + 2K/n\}$. The level constraint and the type-II error exponent can then be verified using the subset and superset, respectively. See Appendix A for details. □

The tests in this section still require closed-form 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 type-II error exponent is indeed achievable, giving a meaningful optimality criterion for goodness-of-fit tests. In the next section, we consider another class of MMD based tests without the need of closed-form integrals.

4.2 Kernel Two-Sample Tests

In the context of model criticism, Lloyd and Ghahramani [35] cast goodness of fit testing into a two-sample problem, where one draws sample y^m from distribution P and then decide if y^m and x^n are from the same distribution. 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 for a two-sample test to attain the error exponent $D(P\|Q)$.

Denote by $\hat{P}_m$ the empirical measure of y^m. Notice that the type-I and type-II error probabilities of a two-sample test depend on both P and Q. We consider the following two-sample test with acceptance region

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

where K is a finite bound on $k(\cdot,\cdot)$,

$$d_k^2(\hat{P}_m, \hat{Q}_n) = \sum_{i=1}^n\sum_{j=1}^n k(x_i, x_j) + \sum_{i=1}^m\sum_{j=1}^m k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^n\sum_{j=1}^m k(x_i, y_j),$$

$$\gamma_{m,n} = \left(\sqrt{K/m} + \sqrt{K/n}\right)\left(2 + \sqrt{-2\log(\alpha/2)}\right).$$

The statistic $d_k^2(\hat{P}_m, \hat{Q}_n)$ for estimating the squared MMD was originally proposed in [25]. Although additional randomness is introduced due to the use of $\hat{P}_m$, it does not hurt the type-II error exponent provided that m is large enough, as stated below.

Theorem 4. *Assume the same conditions as in Theorem 3, 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 rejection region. If m is such that $m/n \to \infty$ as $n \to \infty$, then we have*

1. *under $H_0 : P = Q$, $\mathbf{P}_{y^m x^n}(\Omega_1(m, n)) \leq \alpha$,*
2. *under $H_1 : P \neq Q$,*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbf{P}_{y^m x^n}(\Omega_0(m, n)) = D(P\|Q). \quad (2)$$

The level α constraint can be verified by [25, Theorem 7]. We decompose the type-II error probability into two components and show that each decays at least exponentially at a rate of $D(P\|Q)$. A complete proof is provided in Appendix B.

We may also replace the first two terms in $d_k^2(\hat{P}_m, \hat{Q}_n)$ with $\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i} k(x_i, x_j)$ and $\frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i} k(y_i, y_j)$, which results in an unbiased statistic denoted as $d_u^2(\hat{P}_m, \hat{Q}_n)$ [25]. The following corollary can be shown in a similar manner to Corollary 1 by noting that $|d_u^2(\hat{P}_m, \hat{Q}_n) - d_k^2(\hat{P}_m, \hat{Q}_n)| \leq K/m + K/n$; details are omitted.

Corollary 2. *Under the same assumptions of Theorem 4, the test $d_u^2(\hat{P}_m, \hat{Q}_n) \leq \gamma_{m,n}^2 + K/m + K/n$ has its type-I error probability below α and type-II error exponent being $D(P\|Q)$, when $m/n \to \infty$ as $n \to \infty$.*

4.3 Remarks

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 acceptance threshold [12, 25, 29], making the tests asymptotically level α. These methods, however, introduce additional randomness on the threshold choice and further on the type-II error probability. As a result, it becomes difficult to characterize the type-II error exponent. A simple fix is to take the minimum of the Monte Carlo or bootstrap threshold and the distribution-free one, guaranteeing a vanishing threshold and hence the optimal type-II error exponent. In our experiments, the bootstrap threshold is always smaller than the distribution-free threshold.

Finite vs. Asymptotic Regimes. A finitely positive error exponent $D(P\|Q)$ implies that the error probability decays with $\mathcal{O}\left(2^{-n(D(P\|Q)-\epsilon)}\right)$ where $\epsilon \in (0, D(P\|Q))$ can be arbitrarily small. It further implies that kernels 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 the specific kernel. Selecting a proper kernel is an ongoing research topic and we refer the reader to related works such as [29, 26, 46].

Non-i.i.d. Sample. We notice that Chwialkowski et al. [12] considered non-i.i.d. sample by use of wild bootstrap. In general, statistical optimality with non-i.i.d. sample is difficult to establish even for simple hypothesis testing.

General Two-Sample Problem. Studied in Section 4.2 can be seen as a special case of the two-sample problem where sample sizes scale in different orders, i.e., $m/n \to \infty$ as $n \to \infty$. A direct extension is to consider the more common setting where $0 < \lim_{n\to\infty} m/n < \infty$. For example, an equal number of real and fake samples is typically used for training generative models where the MMD acts as a critic to distinguish between them [33, 20, 32]. However, the current approach is not readily applicable, for lacking an extended version of Sanov's theorem that works with two sample sequences. A naive way may try decomposing the acceptance region $\Omega_0(m,n)$ into $\Omega_0'(m) \times \Omega_0''(n)$ with $\Omega_0'(m)$ and $\Omega_0''(n)$ being respectively decided by y^m and x^n, and then apply Sanov's theorem to each set. Unfortunately, such a decomposition is not possible for the MMD based two-sample tests. We postpone a further investigation until Section 6, after studying the KSD based goodness-of-fit tests in the next section.

5 Kernel Stein Discrepancy Based Goodness-of-Fit Tests

In this section, we investigate the KSD based goodness-of-fit tests recently proposed in [12, 29, 34].

Let $\mathcal{X} = \mathbb{R}^d$. Denote by p and q the density functions (w.r.t. Lebesgue measure) of P and Q, respectively. In [12, 34], the KSD is defined 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 $\|f\|_{\mathcal{H}_k} \le 1$ denotes the unit ball in the RKHS $\mathcal{H}_k$, and $s_p(x) = \nabla_x \log p(x)$ is the score function of $p(x)$. An equivalent expression of the KSD is given by

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

where $h_p(x,y) = s_p^T(x)s_p(y)k(x,y) + s_p^T(y)\nabla_x k(x,y) + s_p^T(x)\nabla_y k(x,y) + \mathrm{trace}(\nabla_{x,y}k(x,y))$. Given sample x^n, we may estimate $d_S^2(P,Q)$ by $d_S^2(P,\hat{Q}_n) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h_p(x_i,x_j)$, which is a degenerate V-statistic under the null hypothesis $H_0 : P = Q$ [12].

With $\mathbf{E}_{x\sim Q}\|\nabla_x \log p(x) - \nabla_x \log q(x)\|^2 \le \infty$ and a C_0-universal kernel [10], $d_S(P,Q) = 0$ if and only if $P = Q$ [12, Theorem 2.2]. A nice property of the KSD is that this result requires only the knowledge of $p(x)$ up to the normalization constant. The KSD has also been shown to be lower bounded in terms of the MMD or the bounded Lipschitz metric (involving some unknown constants) under suitable conditions [23]. This indicates that $d_S(P,P_l) \to 0$ only if $P_l \to P$ weakly, which is important to applying Sanov's theorem in our approach.

Unlike the MMD based test statistics, there does not exist a uniform or distribution-free probabilistic bound on $d_S^2(P,\hat{Q}_n)$. As a result, it is difficult to find a test threshold to meet the fixed level constraint for all sample sizes. To proceed, we relax the level constraint to an asymptotic one, and use the result of [12, Proposition 3.2] which shows that $nd_S^2(P,\hat{Q}_n)$ converges weakly to some distribution under H_0.[2] We assume a fixed α-quantile γ_α of the limiting cumulative distribution function, so that $\lim_{n\to\infty} P(d_S^2(P,\hat{Q}_n) > \gamma_\alpha/n) = \alpha$. Then if γ_n is such that $\gamma_n \to 0$ and $\lim_{n\to\infty} n\gamma_n \to \infty$, e.g., $\gamma_n = \sqrt{1/n}\left(1 + \sqrt{-\log\alpha}\right)$, we get $\gamma_n > \gamma_\alpha/n$ in the limit and thus $\lim_{n\to\infty} P(d_S^2(P,\hat{Q}_n) > \gamma_n) \le \alpha$. Similarly, this threshold choice may be poor in the finite sample regime and we can take the minimum of this threshold and a bootstrap one [2, 13, 31]. Together with the weak convergence properties of the KSD, we have the following result.

Theorem 5. *Let P and Q be distributions defined on $\mathbb{R}^d$, with $0 < D(P\|Q) < \infty$ under the alternative hypothesis. Assume x^n i.i.d. $\sim Q$ and set $\gamma_n = \sqrt{1/n}\left(1 + \sqrt{-\log\alpha}\right)$. It follows that*

1. *if h_p is Lipschitz continuous and $\mathbf{E}_{x\sim Q}h_p(x,x) < \infty$, then under $H_0 : P = Q$,*

$$\lim_{n\to\infty} \mathbf{P}_{x^n}\left(d_S^2(P,\hat{Q}_n) > \gamma_n\right) \le \alpha.$$

2. *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 the*

[2] Chwialkowski et al. [12] assume τ-mixing as the notion of dependence within the observations, which holds in the i.i.d. case. They also assume a technical condition $\sum_{t=1}^\infty t^2\sqrt{\tau(t)} \le \infty$ on τ-mixing. See details in [12, 18].

sequence $\{\hat{Q}_n\}_{n\geq 1}$ *is uniformly tight; 3)* $k(x,y) = (c^2 + \|x-y\|_2^2)^\eta$ *for* $c > 0$ *and* $-1 < \eta < 0$*, then under* $H_1 : P \neq Q$*,*

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

Proof (sketch). The condition for the asymptotic level constraint is taken from [12, Proposition 3.2]. To establish the type-II error exponent, let d_W denote the MMD or the bounded Lipschitz metric, which metrize the weak convergence on $\mathcal{P}$. Under each of the three conditions from [23, Theorems 5, 7, and 8], $d_W(P,\hat{Q}_n) \leq g(d_S(P,\hat{Q}_n))$ where $g(d_S) \to 0$ as $d_S \to 0$. Then there exists γ_n' such that $\{x^n : d_S^2(P,\hat{Q}_n) \leq \gamma_n\} \subset \{x^n : d_W^2(P,\hat{Q}_n) \leq \gamma_n'\}$ and $\gamma_n' \to 0$ as $n \to \infty$. Thus, the type-II error exponent is lower bounded by $D(P\|Q)$, following the same argument of Theorem 3. The upper bound is from Chernoff-Stein lemma which also holds for an asymptotic level constraint. □

Liu et al. [34] proposed an unbiased U-statistic $d_{S(u)}^2(P,\hat{Q}_n) = \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i} h_p(x_i,x_j)$ for estimating $d_S^2(P,Q)$. A similar result holds under an additional assumption on the boundedness of $h_p(\cdot,\cdot)$, using the same argument of Corollary 1.

Corollary 3. *Assume the same conditions as in Theorem 5 and further that* $h_p(\cdot,\cdot) \leq H_p$ *for some* $H_p \in \mathbb{R}^+$*. Then the test* $d_{S(u)}^2(P,\hat{Q}_n) \leq \gamma_n + H_p/n$ *is asymptotically level* α *and achieves the optimal type-II error exponent* $D(P\|Q)$*.*

The Weak Convergence Property. To use Sanov's theorem, we find a superset of probability measures for the equivalent acceptance region, which is required to be closed and to converge (in terms of weak convergence) to P in the large sample limit. Without the weak convergence property, the equivalent acceptance region may contain probability measures that are not close to P, and the minimum KLD over the superset would be hard to obtain. An example can be found in [23, Theorem 6] where the KSDs are driven to zero by sequences of probability measures not converging to P. Consequently, our approach does not establish the optimal type-II error exponent for the linear-time KSD based tests in [29, 34], the linear-time kernel two-sample test in [25], the B-test in [52], and a pseudometric based two-sample test in [14], due to lack of the weak convergence property.

6 General Two-Sample Problem

In this section, we investigate the kernel two-sample tests in a more general setting. As discussed in Section 4.3, the key is to establish an extended Sanov's theorem that is able to handle two sample sequences.

6.1 Extended Sanov's Theorem

We define pairwise weak convergence for probability measures: we say $(P_l, Q_l) \to (P,Q)$ weakly if and only if both $P_l \to P$ and $Q_l \to Q$ weakly. We consider $\mathcal{P}\times\mathcal{P}$ endowed with the topology induced by this pairwise weak convergence. It can be verified that this topology is equivalent to the product topology on $\mathcal{P}\times\mathcal{P}$ where each $\mathcal{P}$ is endowed with the topology of weak convergence. An extended version of Sanov's theorem is stated below.

Theorem 6 (Extended Sanov's Theorem)**.** *Let* $\mathcal{X}$ *be a Polish space,* y^m *i.i.d.* $\sim P$*, and* x^n *i.i.d.* $\sim Q$*. Assume* $0 < \lim_{m,n\to\infty} \frac{m}{m+n} = c < 1$*. Then for a set* $\Gamma \subset \mathcal{P}\times\mathcal{P}$*, it holds that*

$$\begin{aligned}
&\inf_{(R,S)\in \operatorname{int}\Gamma} cD(R\|P) + (1-c)D(S\|Q)\\
\geq\ &\limsup_{m,n\to\infty} -\frac{1}{m+n}\log \mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\\
\geq\ &\liminf_{m,n\to\infty} -\frac{1}{m+n}\log \mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\\
\geq\ &\inf_{(R,S)\in \operatorname{cl}\Gamma} cD(R\|P) + (1-c)D(S\|Q),
\end{aligned}$$

where $\operatorname{int}\Gamma$ *and* $\operatorname{cl}\Gamma$ *denote the interior and closure of* Γ *w.r.t. the pairwise weak convergence, respectively.*

We comment that this extension is not apparent as existing tools, e.g., Cramér theorem [19], used for proving Sanov's theorem can only deal with a single distribution. In Appendix C, we first prove the above result in finite sample space and then extend it to general Polish space, with two simple combinatorial lemmas as prerequisites.

6.2 Exact and Optimal Error Exponent

With the extended Sanov's theorem and a vanishing threshold $\gamma_{m,n}$, we are ready to establish the exponential decay of the type-II error probability. A proof is provided in Appendix D.

Theorem 7. *Assume the same conditions as in Theorem 4, and* $\lim_{m,n\to\infty}\frac{m}{m+n} = c \in (0,1)$*. Under the alternative hypothesis* $H_1 : P \neq Q$*, further assume that*

$$0 < D^* := \inf_{R\in\mathcal{P}} cD(R\|P) + (1-c)D(R\|Q) < \infty.$$

Given $0 < \alpha < 1$*, the test* $d_k(\hat{P}_m,\hat{Q}_n) \leq \gamma_{m,n}$ *with* $\gamma_{m,n}$ *defined in Section 4.2 is level* α *and also exponentially consistent with the type-II error exponent being*

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

Here we consider the error exponent w.r.t. $m+n$, the total number of observations for testing. Therefore, when

$0 < c < 1$, the type-II error probability vanishes as $\mathcal{O}(2^{-(m+n)(D^*-\epsilon)})$, where $\epsilon \in (0, D^*)$ is fixed and can be arbitrarily small. Similarly, this result only requires kernels be bounded continuous and characteristic.

Our next theorem provides an upper bound on the type-II error exponent of any (asymptotically) level α two-sample test. This further shows that the kernel test $d_k(\hat{P}_m, \hat{Q}_n) \leq \gamma_{m,n}$ is asymptotically optimal, by choosing the type-II error exponent as the performance metric. See Appendix E for a proof.

Theorem 8. *Assume the same conditions as in Theorem 7. For a nonparametric two-sample test* $\Omega'(m,n) = \{\Omega'_0(m,n), \Omega'_1(m,n)\}$ *which is (asymptotically) level* $\alpha, 0 < \alpha < 1$*, its type-II error exponent is bounded by* D^**, that is,*

$$\liminf_{m,n\to\infty} -\frac{1}{m+n} \log \mathbf{P}_{y^m x^n}(\Omega'_0(m,n)) \leq D^*.$$

We can use Theorems 7 and 8 to identify more asymptotically optimal two-sample tests:

- Assuming $n = m$, the unbiased test $d_u^2(\hat{P}_m, \hat{Q}_n) \leq (4K/\sqrt{n})\sqrt{\log(\alpha^{-1})}$, with a tighter threshold, is also level α [25]. As $k(\cdot,\cdot)$ is finitely bounded by K, its type-II error probability vanishes exponentially at a rate of $\inf_{R\in\mathcal{P}} \frac{1}{2}D(R\|P) + \frac{1}{2}D(R\|Q)$, which can be shown by the same argument of Corollary 1.

- It is also possible to consider a family of kernels for the test statistic [21, 44]. For a given family κ, the test statistic is $\sup_{k\in\kappa} d_k(\hat{P}_m, \hat{Q}_n)$ which also metrizes weak convergence under suitable conditions, e.g., when κ consists of finitely many Gaussian kernels [44, Theorem 3.2]. If K remains to be an upper bound for all $k \in \kappa$, then comparing $\sup_{k\in\kappa} d_k(\hat{P}_m, \hat{Q}_n)$ with $\gamma_{m,n}$ in Section 4.2 results in an asymptotically optimal level α test.

Fair Alternative. In [38], a notion of fair alternative is proposed when investigating how a two-sample test performs as dimension increases. The idea is to fix $D(P\|Q)$ under the alternative hypothesis for all dimensions, guided by the fact that the KLD is a fundamental information-theoretic quantity determining the hardness of hypothesis testing problems. This approach, however, does not take into account the impact of sample sizes. In light of our results, perhaps a better choice is to fix D^* defined in Theorem 7 when the sample sizes grow in the same order. In practice, D^* may be hard to compute, so fixing its upper bound $(1-c)D(P\|Q)$ and hence $D(P\|Q)$ is reasonable.

Other Discrepancy Measures. Other discrepancy measures between distributions may also metrize the weak convergence on $\mathcal{P}$, including Lévy-Prokhorov metric, the bounded Lipschitz metric, and Wasserstein distance. We may directly compute such a discrepancy between the empirical measures and then compare it with a decreasing threshold. However, there also does not exist a uniform or distribution-free threshold such that the level constraint is satisfied for all sample sizes. A possible remedy, as in Section 5, is to relax the level constraint to an asymptotic one. We will not expand into this direction, as computing such discrepancy measures from samples is generally more costly than the MMD and KSD based statistics.

7 Concluding Remarks

In this paper, we established the statistical optimality of the MMD and KSD based goodness-of-fit tests in the spirit of universal hypothesis testing. The KSD based tests are more computationally efficient, as there is no need to draw samples or compute integrals. In comparison, the MMD based tests are statistically favorable, as they require weaker assumptions and can meet the level constraint for any sample size. The quadratic-time MMD based two-sample tests are also shown to be optimal when sample sizes scale in the same order. Our findings not only solve a long-standing open problem in statistics, but also provide meaningful optimality criteria for nonparametric goodness-of-fit and two-sample testing.

While the optimality criterion is defined in the asymptotic sense, we also conduct experiments of these kernel based goodness-of-fit tests in the finite sample regime, with results given in Appendix F due to space limit. Whereas we cannot tell much statistical difference in our experiments, some experiments in the literature showed that the MMD based tests performed better than the KSD based tests and others showed the opposite [12, 23, 34, 29]. The finite sample performance depends on kernel choice as well as specific distributions. Under the universal setting, no test is known to be optimal in terms of the type-II error probability subject to a given level constraint. Statistical optimality can only be established in the large sample limit, as the one considered in the present work.

Acknowledgement

The authors are grateful to the anonymous reviewers for valuable comments and suggestions. The work of BC was supported in part by the U.S. National Science Foundation under grant CNS-1731237 and by the U.S. Air Force Office of Scientific Research under grant FA9550-16-1-0077. Part of this work was done when SZ and PY were students at Syracuse University.

References

[1] Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006.

[2] M. A. Arcones and E. Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674, 1992.

[3] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2007.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[5] K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *arXiv preprint arxiv:1709.08148*, 2017.

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

[7] 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 (3):309–318, 1994.

[8] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

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

[10] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[11] G. Casella and R. Berger. *Statistical Inference*. Duxbury Thomson Learning, 2002.

[12] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2016.

[13] K. P. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, 2014.

[14] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, 2015.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 2nd edition, 2006.

[16] I. Csiszár. A simple proof of Sanov's theorem. *Bulletin of the Brazilian Mathematical Society*, 37(4):453–459, 2006.

[17] I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

[18] J. Dedecker, P. Doukhan, G. Lang, J. Leon, S. Louhichi, and C. Prieur. *Weak Dependence: With Examples and Applications*. New York: Springer, 2007.

[19] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. New York: Springer, 2009.

[20] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015.

[21] K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in neural information processing systems*, 2009.

[22] J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *NIPS*, 2015.

[23] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *ICML*, 2017.

[24] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, 2009.

[25] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar): 723–773, 2012.

[26] 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 *Advances in Neural Information Processing Systems*, 2012.

[27] L. Györfi and E. C. Van Der Meulen. A consistent goodness of fit test based on the total variation distance. In *Nonparametric Functional Estimation and Related Topics*, pages 631–645. Springer, 1991.

[28] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pages 369–401, 1965.

[29] W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *NIPS*, 2017.

[30] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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

[32] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

[33] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, 2015.

[34] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, 2016.

[35] J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

[36] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[37] 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.

[38] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. A. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, 2015.

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

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

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

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

[43] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, 2007.

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

[45] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

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

[47] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, 2015.

[48] Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17 (1):5272–5311, 2016.

[49] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli. Universal and composite hypothesis testing via mismatched divergence. *IEEE Transactions on Information Theory*, 57(3):1587–1603, 2011.

[50] T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[51] P. Yang and B. Chen. Robust Kullback-Leibler divergence and universal hypothesis testing for continuous distributions. *arxiv preprint arxiv:1711.04238*, 2017.

[52] W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, 2013.

[53] O. Zeitouni and M. Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inf. Theory*, 37(2):285–290, 1991.