

Kernel Robust Hypothesis Testing

Zhongchang Sun and Shaofeng Zou

Abstract—The problem of robust hypothesis testing is studied, where under the null and the alternative hypotheses, the data-generating distributions are assumed to be in some uncertainty sets, and the goal is to design a test that performs well under the worst-case distributions over the uncertainty sets. In this paper, uncertainty sets are constructed in a data-driven manner using kernel method, i.e., they are centered around empirical distributions of training samples from the null and alternative hypotheses, respectively; and are constrained via the distance between kernel mean embeddings of distributions in the reproducing kernel Hilbert space, i.e., maximum mean discrepancy (MMD). The Bayesian setting and the Neyman-Pearson setting are investigated. For the Bayesian setting where the goal is to minimize the worst-case error probability, an optimal test is firstly obtained when the alphabet is finite. When the alphabet is infinite, a tractable approximation is proposed to quantify the worst-case average error probability, and a kernel smoothing method is further applied to design test that generalizes to unseen samples. A direct robust kernel test is also proposed and proved to be exponentially consistent. For the Neyman-Pearson setting, where the goal is to minimize the worst-case probability of miss detection subject to a constraint on the worst-case probability of false alarm, an efficient robust kernel test is proposed and is shown to be asymptotically optimal. Numerical results are provided to demonstrate the performance of the proposed robust tests.

Index Terms—Kernel robust test, Bayesian setting, asymptotic Neyman-Pearson setting, tractable approximation, kernel smoothing

I. INTRODUCTION

HYPOTHESIS testing is a fundamental problem in statistical inference where the goal is to distinguish among different hypotheses with a small probability of error [3]–[5]. The likelihood ratio test is known to be optimal under different settings, e.g., the Neyman-Pearson setting and the Bayesian setting [3], [5]. For example, for binary hypothesis testing, we compare the likelihood ratio between the two hypotheses with a pre-specified threshold to make the decision. Therefore, the data-generating distributions under different hypotheses are needed. In practice, these distributions are usually estimated from historical data or designed using domain knowledge, and thus may deviate from the true data-generating distributions. When the distributions applied in the likelihood ratio test deviate from the true data-generating distributions, the performance of the test may degrade significantly. To address this problem, the approach of robust hypothesis testing is proposed, e.g., [6]–[21], where uncertainty sets are introduced to model the uncertainty in the underlying distributions. Generally, the

uncertainty sets are constructed as collections of distributions that lie in the neighborhood of nominal distributions based on some distance measure. The goal is to design a test that performs well under the worst-case distributions over the uncertainty sets.

The robust hypothesis testing problem has been widely studied and various ways of constructing uncertainty sets have been introduced (see, e.g., [8], [21] for a review). The ϵ -contamination uncertainty sets and the total variation uncertainty sets were investigated in [6] and a censored likelihood ratio test was constructed and shown to be minimax optimal. The problem with uncertainty sets defined via the Kullback-Leibler (KL) divergence was investigated in [7], [8]. The least-favorable distributions (LFDs) were identified under some conditions, and the robust likelihood ratio test based on the LFDs were constructed. In [12], the robust hypothesis testing problem under the Bernoulli distribution was investigated. In [13], the uncertainty sets were constructed via distortion constraints. In those works, nominal distributions are usually estimated from historical data. However, when it comes to the high-dimensional data, which is common in the big data era, it is in general difficult to obtain an accurate estimate of the data-generating distributions. Existing studies are mostly limited to the 1-dimensional case, and a generalization to high-dimensional data, e.g., finding the LFDs, is still an open problem in the literature.

In this paper, we employ a data-driven approach [9], [10] to construct the nominal distributions, and extend the robust hypothesis testing problem to the high-dimensional setting. Specifically, a number of training samples are available from the null and alternative hypotheses, respectively, and their empirical distributions are used as the nominal distributions to design the uncertainty sets. We note that in this case, the uncertainty sets defined via the KL divergence [7], [8] are not applicable, since such uncertainty sets only contain distributions supported on the training samples, which may be problematic if the alphabet is actually infinite.

In [9], [10], the robust hypothesis testing problem was investigated where uncertainty sets are centered around empirical distributions via the Wasserstein distance. In [9], the original 0-1 loss, i.e., the error probability, was firstly smoothed. Then, this relaxed formulation can be solved efficiently, and the LFDs and the nearly-optimal robust detector were identified. In [10], the minimax problem with the 0-1 loss, i.e., the exact probability of error, was considered, where a computationally tractable reformulation and the optimal robust test were characterized. In [11], the data-driven robust hypothesis testing problem with the Sinkhorn distance, which is a variant of Wasserstein distance with entropic regularization, was studied. The original 0-1 loss, i.e., the error probability, was smoothed as in [9]. Then, a finite-dimensional convex optimization

Z. Sun and S. Zou are with the Department of Electrical Engineering, University at Buffalo, Buffalo, NY 14228 USA (e-mail: zhongcha@buffalo.edu, szou3@buffalo.edu).

This paper was presented in part at the 2021 IEEE International Symposium on Information Theory [1] and the 2022 IEEE International Symposium on Information Theory [2].

problem was proposed to approximate the smoothed problem. The solutions were further used to approximate the LFDs, and design the robust test. However, Wasserstein distance based approach has certain drawbacks. First, the Wasserstein distance between the empirical distribution with m samples and its data-generating distribution is bounded by $\mathcal{O}(m^{-1/d})$ [22], which depends on the dimension d of the data. Therefore, when choosing radii of uncertainty sets to guarantee that the true data-generating distributions lie in the uncertainty sets with high probability, it is too pessimistic when d is large. Moreover, coefficients in such a concentration bound depend on the true distribution which is unknown, and thus makes it difficult to use in practice. Second, Wasserstein distance is computationally expensive, especially in the high-dimensional setting.

Moment information, such as mean and variance, is usually used to measure the difference between distributions. In [23], the uncertainty sets are constructed using moment classes, where a finite alphabet was considered, and an asymptotically optimal test was designed. Specifically, the moment uncertainty sets in [23] are defined as $\{P : E_P[f] \leq \theta\}$, where f is a real-valued function, $E_P[f]$ denotes the expectation of f under P , and θ is a constant. In this paper, we generalize the moment classes to the reproducing kernel Hilbert space (RKHS) [24]–[26] and construct uncertainty sets using the maximum mean discrepancy (MMD). Specifically, let $f = g - E_{\hat{P}}[g]$, where \hat{P} is the empirical distribution of samples from P , g is any function in the RKHS. We then consider the worst-case and take the supremum of g with bounded norm over the RKHS. This leads to uncertainty sets centered at \hat{P} and defined by MMD (see more details in Section II). Compared with the Wasserstein distance, the kernel MMD between the empirical distribution with m samples and its data-generating distribution can be bounded by $\mathcal{O}(1/\sqrt{m})$ [27], [28], which is dimension-free and also this bound does not depend on the data-generating distribution. This makes it much easier to choose the radii of the uncertainty sets. Moreover, the kernel MMD is computationally efficient to evaluate.

The MMD-based test statistic has been widely used in statistical signal processing and machine learning. For the one-sample testing problem, where the goal is to distinguish if a sequence of samples come from a certain distribution, and the two-sample testing problem, where the goal is to distinguish if two sequences of samples come from the same distribution, the MMD-based methods and some variants are proposed in [29]–[37]. In [38] and [39], MMD-based tests are proposed to detect anomalous data streams and anomalous network structures, respectively, where the anomalous samples generated from a different distribution from the normal samples. In [40], an MMD-based M-statistic is proposed for data-driven quickest change detection. In our paper, we apply MMD to design robust tests which perform well under the worst-case distributions over the uncertainty sets.

A. Main Contributions

In this paper, we develop a data-driven approach with the kernel method to design uncertainty sets for the problem of

robust hypothesis testing. Specifically, empirical distributions are used directly as the nominal distributions, which avoids the estimation error when fitting the data into a parametric family of distributions. We then use the kernel MMD as the distance metric, which can be viewed as a generalization of the moment classes [23]. The advantage of the kernel method is that it scales well for high-dimensional data, and choosing radii of the uncertainty set does not require the knowledge of the underlying true distribution. More importantly, our designed uncertainty sets contain continuous distributions (not only distributions supported on the training data), and our robust kernel test generalizes with guaranteed out-of-sample performance.

We first focus on the Bayesian setting where the goal is to minimize the worst-case error probability. We first study the case with a finite alphabet, and reformulate the original problem equivalently to a finite-dimensional convex optimization problem via the strong duality of kernel robust optimization [41] and then derive the optimal robust test. For the case with an infinite alphabet, we propose a tractable approximation to quantify the worst-case error probability. The basic idea is to generate a finite number of samples randomly, and reduce the uncertainty set to be supported on these samples. We then rewrite equivalently the original problem as a convex optimization problem, and approximate the original objective function value using the approximated uncertainty set supported on these randomly generated samples. This approximation is tractable since it is a finite-dimensional convex optimization, and it builds connection between the finite-alphabet case and the infinite-alphabet case. We then show that the solutions to the approximation converge almost surely to the solutions to the original infinite-alphabet problem as the number of randomly generated samples goes to infinity. The LFDs (for the approximation) can be recovered, which are also supported on these randomly generated samples. To generalize to unseen data, we further apply the kernel smoothing method on the LFDs, and design a robust test that is the likelihood ratio test between the smoothed LFDs. The computational complexity lies in solving a finite-dimensional convex optimization problem the complexity of which depends on the number of randomly generated samples, and implementing the test using kernel smoothed LFDs the complexity of which is quadratic in the number of randomly generated samples and testing samples. We also propose a direct robust kernel test that can be implemented with a quadratic complexity in the number of samples, and show that it is exponentially consistent. The basic idea is to compare the closest MMD distance between the empirical distribution of samples and the two uncertainty sets.

We then study the Neyman-Pearson setting, where the goal is to minimize the worst-case probability of miss detection subject to a constraint on the worst-case probability of false alarm. We first develop the universal upper bound on the error exponent of miss detection under the Neyman-Pearson setting. The analysis is based on a generalization of the Chernoff-Stein lemma [4], [42]. We then design a novel robust kernel test, which is to compare the closest distance between the empirical distribution of the test samples and the uncertainty set with

a threshold. We further demonstrate that it is asymptotically optimal under the Neyman-Pearson setting. Our proposed robust kernel test does not need to solve for the LFDs, which might be computationally intractable in practice. We also show that our test can be implemented efficiently with a quadratic complexity in the number of samples.

B. Paper Organization

In Section II, we present the preliminaries on MMD and the problem formulation. In Section III, we focus the Bayesian setting, and derive the optimal test for the case with a finite alphabet. For the case with an infinite alphabet, we provide a tractable approximation to quantify the worst-case error probability and propose a kernel smoothing robust test. We also propose an exponentially consistent direct robust kernel test. In Section IV, we study the robust hypothesis testing under the Neyman-Pearson setting, and propose an asymptotically optimal robust kernel test. In Section V, we provide numerical results to validate our theoretical analysis. In Section VI, we present some concluding remarks.

II. PRELIMINARIES AND PROBLEM FORMULATION

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set where samples are taken from. Denote by \mathcal{P} the set of all probability measures on \mathcal{X} .

A. Maximum Mean Discrepancy (MMD)

We first give a brief introduction to the kernel mean embedding and the MMD [24], [25]. Let \mathcal{H} denote the RKHS associated with a kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Specifically, $k(x, \cdot)$ denotes the feature map: $\mathcal{X} \rightarrow \mathcal{H}$, and $k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}}$ defines an inner product on \mathcal{H} . In this paper, we consider the bounded kernel: $0 \leq k(x, x') \leq K$, $\forall x, x' \in \mathcal{X}$, where $K > 0$ is some positive constant. The kernel mean embedding of a distribution is a mapping from \mathcal{P} to \mathcal{H} defined as $\mu_P = \int k(x, \cdot) dP$. Let $E_P[f]$ denote the expectation of a function $f \in \mathcal{H}$. Denote by $\|\cdot\|_{\mathcal{H}}$ the norm on \mathcal{H} . Define the MMD between two distributions P_0 and P_1 as:

$$d_{\text{MMD}}(P_0, P_1) = \sup_{f \in \mathcal{H}: \|f\|_{\mathcal{H}} \leq 1} E_{P_0}[f(x)] - E_{P_1}[f(x)]. \quad (1)$$

With the reproducing property of the RKHS, we have that $E_P[f] = \langle f, \mu_P \rangle_{\mathcal{H}}$. The MMD between P_0 and P_1 can be equivalently written as the distance between μ_{P_0} and μ_{P_1} in the RKHS [26]:

$$\begin{aligned} d_{\text{MMD}}(P_0, P_1) &= \|\mu_{P_0} - \mu_{P_1}\|_{\mathcal{H}} \\ &= \left(E_{x \sim P_0, x' \sim P_0} [k(x, x')] + E_{y \sim P_1, y' \sim P_1} [k(y, y')] \right. \\ &\quad \left. - 2 E_{x \sim P_0, y \sim P_1} [k(x, y)] \right)^{1/2}. \end{aligned} \quad (2)$$

Given samples $x^n = (x_1, x_2, \dots, x_n) \sim P_0$ and $y^m = (y_1, y_2, \dots, y_m) \sim P_1$, an unbiased estimate of the squared MMD [26] between P_0 and P_1 is

$$\begin{aligned} \hat{d}_{\text{MMD}}^2(P_0, P_1) &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(x_i, x_j) \\ &\quad + \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, y_j). \end{aligned} \quad (3)$$

If a kernel k is characteristic [43], the kernel mean embedding is injective, and then $d_{\text{MMD}}(\cdot, \cdot)$ is a metric on \mathcal{P} [26], [44]. In this paper, we consider kernels such that the weak convergence on \mathcal{P} is metrized by MMD [45], [46], e.g., Gaussian kernels and Laplacian kernels.

B. Problem Setup

Let $\mathcal{P}_0, \mathcal{P}_1 \subseteq \mathcal{P}$ denote the uncertainty sets under the null and alternative hypotheses, respectively. We propose a data-driven approach to construct the uncertainty sets. Instead of fitting nominal probability distributions in a parametric form, we have two sequence of training samples: $\hat{x}_0^m = (\hat{x}_{0,1}, \hat{x}_{0,2}, \dots, \hat{x}_{0,m})$ and $\hat{x}_1^m = (\hat{x}_{1,1}, \hat{x}_{1,2}, \dots, \hat{x}_{1,m})$ from the two hypotheses, respectively. Let $\hat{Q}_m^l = \frac{1}{m} \sum_{i=1}^m \delta_{\hat{x}_{l,i}}$ be the empirical distribution of \hat{x}_l^m , $l = 0, 1$, where $\delta_{\hat{x}_{l,i}}$ denotes the Dirac measure on $\hat{x}_{l,i}$. The nominal distributions are then the empirical distributions of data from the two hypotheses, respectively. The uncertainty sets $\mathcal{P}_0, \mathcal{P}_1$ are defined via the MMD:

$$\mathcal{P}_l = \left\{ P \in \mathcal{P} : d_{\text{MMD}}(P, \hat{Q}_m^l) \leq \theta \right\}, \quad l = 0, 1, \quad (4)$$

where θ is the pre-specified radius of the uncertainty sets, and shall be chosen to guarantee that the population distribution falls into the uncertainty sets with high probability. It is assumed that $\mathcal{P}_0, \mathcal{P}_1$ do not overlap, i.e., $\theta < \frac{\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}}{2}$. Otherwise, the problem is trivial.

In [23], the moment class is defined as $\{P \in \mathcal{P} : E_P[f] \leq \theta\}$, where f is real-value function on \mathcal{X} . In the definition of moment class, if we let $f = g - E_{\hat{Q}_m^l}[g]$ and take the supremum over g with $\|g\|_{\mathcal{H}} \leq 1$ in the RKHS, it is then the MMD between P and \hat{Q}_m^l . Therefore, the MMD uncertainty sets can be viewed as a generalization of moment classes to the RKHS.

In this paper, we focus on the robust hypothesis testing problem with MMD uncertainty sets under the Bayesian setting and the Neyman-Pearson setting.

1) *Bayesian Setting.* Given a sample x following an unknown distribution Q , the goal is to distinguish between the null hypothesis $H_0 : Q \in \mathcal{P}_0$ and the alternative hypothesis $H_1 : Q \in \mathcal{P}_1$. For a randomized test $\phi : \mathcal{X} \rightarrow [0, 1]$, it accepts the null hypothesis H_0 with probability $1 - \phi(x)$ and accepts the alternative hypothesis H_1 with probability $\phi(x)$. Let

$$\begin{aligned} P_F(\phi) &\triangleq \sup_{P_0 \in \mathcal{P}_0} E_{P_0}[\phi(x)], \\ P_M(\phi) &\triangleq \sup_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \phi(x)] \end{aligned} \quad (5)$$

denote the worst-case probability of false alarm (type-I error probability) and the worst-case probability of miss detection (type-II error probability) for the test ϕ .

For the simple hypothesis testing with equal priors on the two hypotheses, the error probability in the Bayesian setting is given by

$$P_E(\phi) \triangleq \frac{1}{2}E_{P_0}[\phi(x)] + \frac{1}{2}E_{P_1}[1 - \phi(x)] \\ = \frac{1}{2} \int \phi(x)dP_0(x) + \frac{1}{2} \int (1 - \phi(x))dP_1(x), \quad (6)$$

where P_0 and P_1 denote the distributions under the null and alternative hypotheses, respectively. For the Bayesian robust hypothesis testing, the goal is to solve the following problem:

$$\inf_{\phi} \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} P_E(\phi). \quad (7)$$

The results in this paper can be easily generalized to the case with non-equal priors.

Denote a sequence of independent and identically distributed (i.i.d.) samples by $x^n = (x_1, x_2, \dots, x_n)$. The worst-case type-I error exponent e_F and the worst-case type-II error exponent e_M are defined as follows:

$$e_F(\phi) = \inf_{P_0 \in \mathcal{P}_0} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_0}[\phi(x^n)], \\ e_M(\phi) = \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)]. \quad (8)$$

Definition 1. A test ϕ is said to be exponentially consistent if $e_F(\phi) > 0$ and $e_M(\phi) > 0$.

2) *Neyman-Pearson Setting.* In this paper, we focus on the asymptotic Neyman-Pearson setting, where the goal is to solve the following problem:

$$\sup_{\phi: P_F(\phi) \leq \alpha} \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)], \quad (9)$$

where $\alpha \in (0, 1]$ is a pre-specified constraint on the worst-case false alarm probability. Specifically, among the tests that satisfy the false alarm constraint $P_F(\phi) \leq \alpha$, we aim to find one that maximizes the worst-case type-II error exponent.

In this paper, any distributions P_0, P_1 are assumed to admit probability density functions (PDFs) p_0, p_1 , since we can always choose a reference measure μ such that both P_0 and P_1 are absolutely continuous with respect to μ . In general, μ can be chosen as $\mu = P_0 + P_1$. For the continuous distributions and the discrete distributions, μ can be chosen as the Lebesgue measure and the counting measure, respectively.

III. ROBUST HYPOTHESIS TESTING UNDER BAYESIAN SETTING

In this section, we focus on the Bayesian setting. We aim to solve the minimax problem for the average probability of error in (7).

A. Finite-Alphabet Case

Consider the case with a finite alphabet, i.e., $N \triangleq |\mathcal{X}| < \infty$. Let $\mathcal{X} = \{z_1, z_2, \dots, z_N\}$. Then, $\hat{x}_{l,j} \in \{z_i\}_{i=1}^N$ for $l = 0, 1, j = 1, \dots, m$. In this case,

$$P_E(\phi) = \sum_{i=1}^N (1 - \phi_N(z_i))P_1^N(z_i) + \phi_N(z_i)P_0^N(z_i), \quad (10)$$

where we introduce the superscript N on P_0 and P_1 to emphasize its dependence on N . Therefore, (7) can be written as

$$\inf_{\phi_N \in [0,1]^{\otimes N}} \sup_{P_0^N \in \mathcal{P}_0, P_1^N \in \mathcal{P}_1} \sum_{i=1}^N ((1 - \phi_N(z_i))P_1^N(z_i) + \phi_N(z_i)P_0^N(z_i)). \quad (11)$$

Note that (11) is a minimax problem. We then provide the following strong duality result for (11), which is a finite-dimensional convex optimization problem.

Lemma 1. The minimax problem in (11) has the following strong dual formulation:

$$\inf_{\phi_N \in [0,1]^{\otimes N}, f_0, g_0, \alpha_j, \beta_j \in \mathbb{R}} f_0 + g_0 + \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^N \alpha_j k(z_j, \hat{x}_{1,i}) \\ + \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^N \beta_j k(z_j, \hat{x}_{0,i}) + \theta \left\| \sum_{j=1}^N \alpha_j k(z_j, \cdot) \right\|_{\mathcal{H}} \\ + \theta \left\| \sum_{j=1}^N \beta_j k(z_j, \cdot) \right\|_{\mathcal{H}} \\ \text{subject to } 1 - \phi_N(z_i) \leq f_0 + \sum_{j=1}^N \alpha_j k(z_j, z_i), i = 1, \dots, N \\ \phi_N(z_i) \leq g_0 + \sum_{j=1}^N \beta_j k(z_j, z_i), i = 1, \dots, N \\ 0 \leq \phi_N(z_i) \leq 1, i = 1, \dots, N, \quad (12)$$

which is a finite-dimensional convex optimization problem.

Proof. From the strong duality of kernel robust optimization [41, Theorem 3.1], we have that (11) has the following dual problem and the strong duality holds.

$$\inf_{\phi_N \in [0,1]^{\otimes N}, f_0, g_0 \in \mathbb{R}, f_1, g_1 \in \mathcal{H}} f_0 + g_0 + \frac{1}{m} \sum_{i=1}^m f_1(\hat{x}_{1,i}) + \frac{1}{m} \sum_{i=1}^m g_1(\hat{x}_{0,i}) \\ + \theta \|f_1\|_{\mathcal{H}} + \theta \|g_1\|_{\mathcal{H}} \\ \text{subject to } 1 - \phi_N(z_i) \leq f_0 + f_1(z_i), i = 1, \dots, N \\ \phi_N(z_i) \leq g_0 + g_1(z_i), i = 1, \dots, N \\ 0 \leq \phi_N(z_i) \leq 1, i = 1, \dots, N, \quad (13)$$

From the robust representer theorem [41], the functions f_1, g_1 admit the finite expansions $f_1(\cdot) = \sum_{j=1}^N \alpha_j k(z_j, \cdot)$ and $g_1(\cdot) = \sum_{j=1}^N \beta_j k(z_j, \cdot)$. Therefore, the optimization problem in (11) can be reformulated as a finite-dimensional convex optimization problem thus can be solved efficiently in practice. \square

Note that (12) is a convex optimization problem with linear constraints, and thus can be solved using standard optimization tools [47]. By solving (12), we obtain the optimal robust test ϕ_N^* and can also find the optimal solutions $P_0^{*,N}, P_1^{*,N}$ for the inner problem in (11) by plugging ϕ_N^* back to (11). When ϕ_N^* is known, (11) reduces to a finite-dimensional convex optimization problem and can be solved efficiently. In the following section, we also show that the results in the finite-alphabet case can be used to provide an asymptotically accurate approximation for the infinite-alphabet case.

B. Infinite-Alphabet Case

Consider the case where \mathcal{X} is infinite. Then, (12) is infinite-dimensional, and is not directly solvable. To simplify the analysis of (7), we first interchange the sup and inf operators in (7) based on the following proposition. Since the likelihood ratio test is optimal for the binary hypothesis testing problem, the inner problem can be solved by applying the likelihood ratio test. The original problem is then converted to a maximization problem.

Proposition 1. *The minimax problem in (7) has the following reformulation:*

$$\begin{aligned} \inf_{\phi} \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} P_E(\phi) &= \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \inf_{\phi} P_E(\phi) \\ &= \frac{1}{2} \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \int \min \{p_0(x), p_1(x)\} dx. \end{aligned} \quad (14)$$

Proof. The error probability $P_E(\phi)$ is continuous, real-valued and linear in ϕ , P_0 and P_1 . For any distributions $Q_1, Q_2 \in \mathcal{P}_l$, $l = 0, 1$, from the triangle inequality of MMD [26], the convex combination $\lambda Q_1 + (1 - \lambda)Q_2$, $0 < \lambda < 1$, lies in \mathcal{P}_l . Therefore, the uncertainty set \mathcal{P}_0 and \mathcal{P}_1 are convex sets and $\mathcal{P}_0 \times \mathcal{P}_1$ is also convex. Denote by Φ the collection of all ϕ . We have that Φ is the product of uncountably many compact sets of $[0, 1]$. Since \mathcal{X} is compact, from the Tychonoff's theorem [48], [49], Φ is compact with respect to the product topology. Moreover, for any $\phi_1, \phi_2 \in \Phi$, the convex combination $\lambda \phi_1 + (1 - \lambda)\phi_2$, $0 < \lambda < 1$, also lies in Φ . Therefore, Φ is convex. From the Sion's minimax theorem [50], we have that

$$\begin{aligned} \inf_{\phi} \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} P_E(\phi) &= \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \inf_{\phi} P_E(\phi) \\ &= \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \frac{1}{2} \int \mathbb{I}_{\left\{\frac{p_1(x)}{p_0(x)} \geq 1\right\}} p_0(x) dx \\ &\quad + \frac{1}{2} \int \mathbb{I}_{\left\{\frac{p_1(x)}{p_0(x)} < 1\right\}} p_1(x) dx \\ &= \frac{1}{2} \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \int \min \{p_0(x), p_1(x)\} dx, \end{aligned} \quad (15)$$

where \mathbb{I} denotes the indicator function and the second equality is due to the fact that the likelihood ratio test is optimal for the binary hypothesis testing problem [3], [5]. \square

Observe that the problem in (14) is an infinite-dimensional optimization problem and the closed-form optimal solutions P_0^*, P_1^* are difficult to derive. In the following, we propose a

tractable approximation for the minimax error probability in (14). With this approximation, the worst-case error probability in (14) can be quantified. The optimal solutions of this tractable approximation can be further used to design a robust test that generalizes to unseen samples.

Let P be an arbitrary distribution supported on the whole space \mathcal{X} , and is absolutely continuous w.r.t. a uniform distribution on \mathcal{X} . Let $\{z_i\}_{i=1}^N$ be N i.i.d. samples generated from P . We then propose the following approximation of (14) by restricting to distributions supported on the N samples:

$$\frac{1}{2} \sup_{P_0^N \in \mathcal{P}_0^N, P_1^N \in \mathcal{P}_1^N} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\}, \quad (16)$$

where \mathcal{P}_l^N ($l = 0, 1$) denotes the collection of distributions that are supported on $\{z_i\}_{i=1}^N$ and satisfy $\|\mu_{P_l^N} - \mu_{Q_m^l}\|_{\mathcal{H}} \leq \theta$. We note that (16) is a finite-dimensional convex optimization problem which can be solved by standard optimization tools. Let

$$\begin{aligned} f(\mathcal{P}_0, \mathcal{P}_1) &= \frac{1}{2} \sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \int \min \{p_0(x), p_1(x)\} dx, \\ f(\mathcal{P}_0^N, \mathcal{P}_1^N) &= \frac{1}{2} \sup_{P_0^N \in \mathcal{P}_0^N, P_1^N \in \mathcal{P}_1^N} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\}. \end{aligned} \quad (17)$$

Clearly, (16) is a lower bound of (14), i.e., $f(\mathcal{P}_0^N, \mathcal{P}_1^N) \leq f(\mathcal{P}_0, \mathcal{P}_1)$. The following theorem demonstrates that as $N \rightarrow \infty$, the value of (16) converges to the value of (14) almost surely.

Theorem 1. *As $N \rightarrow \infty$, $f(\mathcal{P}_0^N, \mathcal{P}_1^N)$ converges to $f(\mathcal{P}_0, \mathcal{P}_1)$ almost surely.*

Before we prove Theorem 1, we will first show that $\int \min \{p_0(x), p_1(x)\} dx$ is upper semi-continuous in P_0, P_1 with respect to the weak convergence in the following lemma.

Lemma 2. *$\int \min \{p_0(x), p_1(x)\} dx$ is upper semi-continuous in P_0, P_1 with respect to the weak convergence.*

Proof. The proof of Lemma 2 can be found in Appendix C. \square

With Lemma 2, we are ready to prove Theorem 1.

Proof sketch. We first show that the solutions to $\sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \int \min \{p_0(x), p_1(x)\} dx$ exist and let P_0^*, P_1^* denote the optimal solutions. We then show that there exist distributions P_0^s, P_1^s supported on s samples converging weakly to P_0^*, P_1^* respectively, as $s \rightarrow \infty$. Thirdly, for a fixed s , we show that there exist distributions $P_0^{s,N}, P_1^{s,N}$ supported on $\{z_i\}_{i=1}^N$ converging weakly to P_0^s, P_1^s almost surely as $N \rightarrow \infty$ and

$$\sum_{i=1}^N \min \{P_0^{s,N}(z_i), P_1^{s,N}(z_i)\} \geq \int \min \{p_0^*(x), p_1^*(x)\} dx. \quad (18)$$

Moreover, we show that for any $\epsilon > 0$, there exists large s and N such that the MMD between $P_0^{s,N}, P_1^{s,N}$ and \hat{Q}_m^0, \hat{Q}_m^1

can be bounded by $\theta + \epsilon$. Finally, by letting $\epsilon \rightarrow 0$, we prove the convergence result in Theorem 1.

The full proof of Theorem 1 can be found in Appendix D. \square

Though the optimal solutions P_0^*, P_1^* for (14) is difficult to derive, (16) provides a lower bound on the worst-case error probability, and is asymptotically accurate as $N \rightarrow \infty$.

C. Convergence Rate

In this section, we characterize the approximation error in Theorem 1 in terms of N if radial basis function (RBF) kernels [51] are used, i.e., $k(x, y) = \exp(-\frac{\|x-y\|_2^2}{2\sigma^2})$, where $\rho > 0$ is some constant. For example, when $\rho = 1$, $k(x, y)$ is the Laplacian kernel (exponential kernel); and when $\rho = 2$, $k(x, y)$ is the Gaussian kernel. A δ -net of \mathcal{X} is a set of points $\{z_i\}_{i=1}^N$ in \mathcal{X} such that for any $z \in \mathcal{X}$, there exists some z_i that satisfies $\|z - z_i\|_2^\rho \leq \delta$. From classic covering number results, we can construct a δ -net where $\delta = \frac{\max_{z, z' \in \mathcal{X}} \|z - z'\|_2^\rho}{N^{\frac{1}{d}}}$ with the number of N points. We use $\{z_i\}_{i=1}^N$ to construct the support sample set. There exists a partition $\mathcal{A}_N = \{\mathcal{A}_N^1, \mathcal{A}_N^2, \dots, \mathcal{A}_N^N\}$ based on the δ -net such that $z_i \in \mathcal{A}_N^i$ and $\max_{z \in \mathcal{A}_N^i} \|z - z_i\|_2^\rho \leq \delta$. We rewrite $f(P_0, P_1), f(P_0^N, P_1^N)$ in (17) as a function of θ and define

$$g(\theta) = \sup_{\substack{P_0 \in \mathcal{P}: \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta \\ P_1 \in \mathcal{P}: \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta}} \int \min\{p_0(x), p_1(x)\} dx, \quad (19)$$

and

$$g_N(\theta) = \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min\{P_0^N(z_i), P_1^N(z_i)\}. \quad (20)$$

We note that for RBF kernels, $0 \leq k(x, y) \leq 1$. Therefore, for any $P_0 \in \mathcal{P}_0$, we have that $\|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} = \left(E_{x \sim P_0, x' \sim P_0} [k(x, x')] + E_{y \sim \hat{Q}_m^0, y' \sim \hat{Q}_m^0} [k(y, y')] - 2E_{x \sim P_0, y \sim \hat{Q}_m^0} [k(x, y)] \right)^{1/2} \leq \sqrt{2}$. Similarly, we have that $\|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \sqrt{2}$ for any $P_1 \in \mathcal{P}_1$. Therefore, it suffices to consider $\theta \in (0, \sqrt{2}]$. We then have the following theorem that bounds the approximation error.

Theorem 2. Let $\epsilon = \sqrt{2 - 2 \exp\left(-\frac{\delta}{2\sigma^2}\right)}$. For any $\theta \in (0, \sqrt{2}]$, the approximation error satisfies $|g(\theta) - g_N(\theta)| \leq L\epsilon$, where L is some constant.

Proof. Denote by P_0^N, P_1^N discrete distributions with $P_0^N(z_i) = P_0^*(\mathcal{A}_N^i), P_1^N(z_i) = P_1^*(\mathcal{A}_N^i)$. Consider the MMD between P_0^N and P_0^* , we then have that

$$\begin{aligned} & \|\mu_{P_0^N} - \mu_{P_0^*}\|_{\mathcal{H}} \\ &= \sup_{h: \|h\|_{\mathcal{H}} \leq 1} \int h dP_0^N - \int h dP_0^* \\ &= \sup_{h: \|h\|_{\mathcal{H}} \leq 1} \sum_{i=1}^N \int_{\mathcal{A}_N^i} (h - h(z_i)) dP_0^* \\ &\leq \sup_{h: \|h\|_{\mathcal{H}} \leq 1} \sum_{i=1}^N \int_{\mathcal{A}_N^i} \max_{z \in \mathcal{A}_N^i} (h(z) - h(z_i)) dP_0^* \\ &\stackrel{(a)}{=} \sup_{h: \|h\|_{\mathcal{H}} \leq 1} \sum_{i=1}^N \int_{\mathcal{A}_N^i} \max_{z \in \mathcal{A}_N^i} \langle h, k(z, \cdot) - k(z_i, \cdot) \rangle_{\mathcal{H}} dP_0^* \\ &\stackrel{(b)}{\leq} \sup_{h: \|h\|_{\mathcal{H}} \leq 1} \sum_{i=1}^N \int_{\mathcal{A}_N^i} \max_{z \in \mathcal{A}_N^i} \|h\|_{\mathcal{H}} \|k(z, \cdot) - k(z_i, \cdot)\|_{\mathcal{H}} dP_0^* \\ &\leq \sup_{h: \|h\|_{\mathcal{H}} \leq 1} \sum_{i=1}^N \int_{\mathcal{A}_N^i} \|h\|_{\mathcal{H}} \\ &\quad \times \max_{z \in \mathcal{A}_N^i} \sqrt{k(z, z) + k(z_i, z_i) - 2k(z, z_i)} dP_0^* \\ &= \sum_{i=1}^N \int_{\mathcal{A}_N^i} \max_{z \in \mathcal{A}_N^i} \sqrt{k(z, z) + k(z_i, z_i) - 2k(z, z_i)} dP_0^* \\ &= \sum_{i=1}^N \int_{\mathcal{A}_N^i} \sqrt{2 - 2 \min_{z \in \mathcal{A}_N^i} k(z, z_i)} dP_0^* \\ &\stackrel{(c)}{\leq} \sqrt{2 - 2 \exp\left(-\frac{\delta}{2\sigma^2}\right)}, \end{aligned} \quad (21)$$

where (a) is from the reproducing property of the kernel, (b) is from the Cauchy-Schwartz inequality and (c) is due to the fact that $\min_{z \in \mathcal{A}_N^i} k(z, z_i) \geq \exp\left(-\frac{\delta}{2\sigma^2}\right)$. Let $\epsilon = \sqrt{2 - 2 \exp\left(-\frac{\delta}{2\sigma^2}\right)}$. We then have that $\|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon$ by the triangle inequality. Therefore, P_0^N lies in the uncertainty set centered around \hat{Q}_m^0 with radius $\theta + \epsilon$. Similarly, for P_1^N and P_1^* , we have the same result that P_1^N lies in the uncertainty set centered around \hat{Q}_m^1 with radius $\theta + \epsilon$. From Jensen's inequality [52], we have that

$$\int \min\{p_0^*(x), p_1^*(x)\} dx \leq \sum_{i=1}^N \min\{P_0^N(z_i), P_1^N(z_i)\}. \quad (22)$$

Therefore, $g_N(\theta + \epsilon) \geq g(\theta)$. We then have that

$$\begin{aligned} |g(\theta) - g_N(\theta)| &\leq |g_N(\theta + \epsilon) - g(\theta)| + |g_N(\theta + \epsilon) - g_N(\theta)| \\ &\leq |g(\theta + \epsilon) - g(\theta)| + |g_N(\theta + \epsilon) - g_N(\theta)|, \end{aligned} \quad (23)$$

where the first inequality is due to the fact that $g_N(\theta + \epsilon) \geq g(\theta)$ and $g(\theta + \epsilon) \geq g_N(\theta + \epsilon)$. We will then bound the first term $|g(\theta + \epsilon) - g(\theta)|$. Since $g(\theta)$ is concave when $\theta \in (0, \infty)$ (see (76) in Appendix D for the proof), there exists L_1 such that $g(\cdot)$ is L_1 -Lipschitz on $[\theta, \sqrt{2}]$ [53]. Therefore, $|g(\theta + \epsilon) - g(\theta)| \leq L_1\epsilon$. Similarly, there exists L_2 such that $|g_N(\theta + \epsilon) -$

$g_N(\theta) \leq L_2\epsilon$. Therefore, $|g(\theta) - g_N(\theta)| \leq (L_1 + L_2)\epsilon$ for any $\theta \in (0, \sqrt{2}]$. Note that L_1 and L_2 may depend on θ . \square

This result characterizes the convergence rate of the approximation error with respect to the support sample size N . In practice, we can choose a proper N to control the approximation error. This result also reveals the relation between the data dimension d and the number of support sample N . In high dimensional setting, we need a larger N to achieve the same approximation error as in the low dimensional setting. Note that N is the number of artificially generated samples, and therefore we could generate as many as we like at the price of increased computational cost for solving (16).

D. Robust Test via Kernel Smoothing

The optimal solutions P_0^*, P_1^* of (14), and thus the likelihood ratio test between P_0^* and P_1^* are difficult to derive. Note that $P_0^{*,N}, P_1^{*,N}$ are optimal solutions to (16). The following proposition shows that the sequence $\{P_0^{*,N}, P_1^{*,N}\}_{N=1}^\infty$ converges weakly to an optimal solution of (14), i.e., for all bounded and continuous functions h , $\lim_{N \rightarrow \infty} E_{P_0^{*,N}}[h] = E_{P_0^*}[h]$ and $\lim_{N \rightarrow \infty} E_{P_1^{*,N}}[h] = E_{P_1^*}[h]$. The fact that $P_0^{*,N}, P_1^{*,N}$ are reasonable approximations of P_0^*, P_1^* as $N \rightarrow \infty$ further motivates our kernel smoothing method to design a robust test that generalizes to the entire alphabet in this section.

Proposition 2. *The sequence $\{P_0^{*,N}, P_1^{*,N}\}_{N=1}^\infty$ converges weakly to an optimal solution of (14).*

Proof. Observe that for any N , $\{P_0^{*,N}, P_1^{*,N}\}$ lies in the compact set $\mathcal{P}_0 \times \mathcal{P}_1$. Assume $\{P_0^{*,N}, P_1^{*,N}\}$ does not converge weakly to the optimal solution $\{P_0^*, P_1^*\}$, then there exists a subsequence of $\{P_0^{*,N}, P_1^{*,N}\}_{N=1}^\infty$ converges weakly but not to $\{P_0^*, P_1^*\}$ [54, Chapter 5.1.1]. Denote the sequence by $\{P_0^{*,N(t)}, P_1^{*,N(t)}\}_{t=1}^\infty$. Assume $\{P_0^{*,N(t)}, P_1^{*,N(t)}\}_{t=1}^\infty$ converges weakly to $\{P_0', P_1'\}$. Since $\{P_0', P_1'\}$ is not an optimal solution to (14), we have that

$$\int \min\{p_0'(x), p_1'(x)\} dx < \int \min\{p_0^*(x), p_1^*(x)\} dx. \quad (24)$$

We then have that

$$\begin{aligned} & \int \min\{p_0^*(x), p_1^*(x)\} dx \\ &= \lim_{t \rightarrow \infty} \sum_{i=1}^N \min\{P_0^{*,N(t)}(z_i), P_1^{*,N(t)}(z_i)\} \\ &\leq \int \min\{p_0'(x), p_1'(x)\} dx, \end{aligned} \quad (25)$$

where the equality is from Theorem 1 and the inequality is due to the upper semi-continuity of $\int \min\{p_0(x), p_1(x)\} dx$ in Lemma 2. This leads to a contradiction. Therefore, $\{P_0^{*,N}, P_1^{*,N}\}_{N=1}^\infty$ converges weakly an optimal solution of (14). \square

Note that $P_0^{*,N}, P_1^{*,N}$ are convex combinations of Dirac measures. We then extend them to the whole space via kernel smoothing to approximate P_0^*, P_1^* , i.e.,

$$\begin{aligned} \tilde{P}_0^*(x) &= \sum_{i=1}^N P_0^{*,N}(z_i) k(x, z_i), \\ \tilde{P}_1^*(x) &= \sum_{i=1}^N P_1^{*,N}(z_i) k(x, z_i). \end{aligned} \quad (26)$$

The kernel functions have various choices. For example, the Gaussian kernel with bandwidth parameter σ : $k(x, y) = \frac{1}{\sqrt{2\sigma}} \exp(-\frac{\|x-y\|^2}{2\sigma^2})$. After kernel smoothing, we define the likelihood ratio test $\tilde{\phi}$ between $\tilde{P}_1^*(x)$ and $\tilde{P}_0^*(x)$ over the whole space \mathcal{X} as follows to approximate the optimal test:

$$\tilde{\phi}(x) = \begin{cases} 1, & \text{if } \log \frac{\tilde{P}_1^*(x)}{\tilde{P}_0^*(x)} \geq 0 \\ 0, & \text{if } \log \frac{\tilde{P}_1^*(x)}{\tilde{P}_0^*(x)} < 0. \end{cases} \quad (27)$$

When the testing sample size is n , after solving $P_0^{*,N}, P_1^{*,N}$, the computational complexity for implementing $\tilde{\phi}$ is $\mathcal{O}(nN)$. The numerical results in Section V show that $\tilde{\phi}$ performs well in practice, and is robust to model uncertainty.

E. A Direct Robust Kernel Test

In this section, we consider the problem of testing a sequence of samples x^n , where n is the sample size. We propose a direct robust kernel test and further show that it is exponentially consistent as $n \rightarrow \infty$ under the Bayesian setting.

Motivated by the facts that the MMD can be used to measure the distance between distributions when the kernel k is characteristic, we propose a direct robust kernel test as follows

$$\phi_B(x^n) = \begin{cases} 1, & \text{if } S(x^n) \geq \gamma \\ 0, & \text{if } S(x^n) < \gamma, \end{cases} \quad (28)$$

where

$$S(x^n) = \inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}}, \quad (29)$$

and γ is a pre-specified threshold. In the construction of $S(x^n)$, we use “inf” to tackle the uncertainty of distributions and compare the closest distance between the empirical distribution of samples and the two uncertainty sets. The test statistic involves two infinite-dimensional optimization problems, and thus is difficult to solve in general. In the following proposition, we show that it can actually be solved analytically in closed-form, and the computational complexity is $\mathcal{O}(m^2 + n^2)$.

Proposition 3. *For any $l = 0, 1$, if $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^l}\|_{\mathcal{H}} > \theta$, then*

$$\begin{aligned} & \inf_{P \in \mathcal{P}_l} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} = \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^l}\|_{\mathcal{H}} - \theta \\ &= \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(\hat{x}_{l,i}, \hat{x}_{l,j}) \right. \\ & \quad \left. - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, \hat{x}_{l,j}) \right)^{1/2} - \theta; \end{aligned} \quad (30)$$

and otherwise, $\inf_{P \in \mathcal{P}_l} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} = 0$.

Proof. The proof of Proposition 3 can be found in Appendix E. \square

In the following theorem, we show that with a proper choice of γ , ϕ_B is exponentially consistent.

Theorem 3. 1) If $\gamma \in \left(-\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + 2\theta, \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \right)$, ϕ_B is exponentially consistent.
2) ϕ_B can be equivalently written as

$$\phi'_B(x^n) = \begin{cases} 1, & \text{if } \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \geq \gamma \\ 0, & \text{if } \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} < \gamma, \end{cases} \quad (31)$$

and its computational complexity is $\mathcal{O}(m^2 + n^2)$.

Proof. The proof of Theorem 3 can be found in Appendix F. \square

It can be seen that the direct test for robust hypothesis testing naturally reduces to comparing the MMD distance between the empirical distribution of samples and two centers of uncertainty sets which is computationally efficient. The exponential consistency of ϕ_B implies that the error probabilities decay exponentially fast with the sample size n . In practice, we can choose a proper threshold to balance the trade-off between the two types of errors.

The error exponent in Theorem 3 is in an asymptotic sense, and is in the form of an optimization problem without a closed-form solution. In the following proposition, we consider a special case with $\gamma = 0$ and derive the closed-form upper bound of the worst-case error probabilities.

Proposition 4. Set $\gamma = 0$ in (28). Then, the worst-case type-I and type-II errors can be bounded as follows,

$$\begin{aligned} & \sup_{P_0 \in \mathcal{P}_0} E_{P_0}[\phi_B(x^n)] \\ & \leq \exp \left(- \frac{n \left(\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 - 2\theta \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \right)^2}{8K^2} \right) \end{aligned} \quad (32)$$

and

$$\begin{aligned} & \sup_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \phi_B(x^n)] \\ & \leq \exp \left(- \frac{n \left(\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 - 2\theta \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \right)^2}{8K^2} \right). \end{aligned} \quad (33)$$

Proof. The proof of Proposition 4 can be found in Appendix G. \square

In Proposition 4, we provide an upper bound on the worst-case error probability of ϕ_B when $\gamma = 0$. It can be seen that the error probabilities decay exponentially fast with an exponent of $\left(\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 - 2\theta \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \right)^2 / 8K^2$, which validates the fact that ϕ_B is exponentially consistent. Moreover, the decay rate is a function of the radius θ and

the MMD distance between centers of two uncertainty sets. When the centers of two uncertainty sets are fixed, the upper bound on the error probabilities will increase with the radius θ . Proposition 4 provides a closed-form non-asymptotic upper bound on the worst-case error probability. In practice, this upper bound can be used to evaluate the worst-case risk of implementing ϕ_B for a finite sample size n . Moreover, combining Theorem 1 and Theorem 3, the performance gap between ϕ_B and the optimal test can be approximated.

IV. ROBUST HYPOTHESIS TESTING UNDER NEYMAN-PEARSON SETTING

In this section, we focus on the Neyman-Pearson setting. We propose a robust kernel test, and show that it is asymptotically optimal under the Neyman-Pearson setting. The results in this section also hold for $\mathcal{X} = \mathbb{R}^d$.

A. Universal Upper Bound on the Worst-Case Error Exponent

In this section, we derive the universal upper bound on the error exponent for the problem in (9). The following proposition is a robust version of the Chernoff-Stein lemma [4], [42].

Proposition 5. Consider the robust hypothesis testing problem in (9), we have that

$$\begin{aligned} & \sup_{\phi: P_F(\phi) \leq \alpha} \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)] \\ & \leq \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1). \end{aligned} \quad (34)$$

Proof. For any $P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1$, from the Chernoff-Stein lemma [4], [42], we have that

$$\sup_{\phi: E_{P_0}[\phi(x^n)] \leq \alpha} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)] = D(P_0 \| P_1). \quad (35)$$

Since $P_F(\phi) \triangleq \sup_{P_0 \in \mathcal{P}_0} E_{P_0}[\phi(x^n)]$, we have that $\{\phi : P_F(\phi) \leq \alpha\} \subseteq \{\phi : E_{P_0}[\phi(x^n)] \leq \alpha\}$. Therefore, for any $P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1$, we have that

$$\begin{aligned} & \sup_{\phi: P_F(\phi) \leq \alpha} \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)] \\ & \leq \sup_{\phi: E_{P_0}[\phi(x^n)] \leq \alpha} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)] \\ & = D(P_0 \| P_1). \end{aligned} \quad (36)$$

The solutions to $\inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1)$ exist since $D(P_0 \| P_1)$ is lower semi-continuous and lower semi-continuous functions attain its infimum on a compact set. Since (36) holds for any $P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1$, we then have that

$$\begin{aligned} & \sup_{\phi: P_F(\phi) \leq \alpha} \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)] \\ & \leq \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1). \end{aligned} \quad (37)$$

\square

Proposition 5 implies that for any test, the achievable error exponent is no better than $\inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1)$. This theorem also applies to robust hypothesis testing problems with different uncertainty sets.

B. Asymptotically Optimal Robust Kernel Test

In this section, we propose a robust kernel test for the problem in (9), and further prove that it is asymptotically optimal.

Motivated by the fact that when the kernel k is characteristic, the MMD is a metric and can be used to measure the distance between distributions and the kernel test is asymptotically optimal under the Neyman-Pearson setting for the two-sample test problem [35], we design our robust kernel test as follows:

$$\phi_N(x^n) = \begin{cases} 1, & \text{if } \inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} > \gamma_n \\ 0, & \text{if } \inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \leq \gamma_n, \end{cases} \quad (38)$$

where $\gamma_n = \sqrt{2K/n}(1 + \sqrt{-\log \alpha})$ is chosen to satisfy the false alarm constraint that $P_F(\phi_N) \leq \alpha$, and \hat{P}_n is the empirical distribution of x^n . We use “inf” in the test statistic to tackle the uncertainty of distributions. A heuristic explanation of our test is that we use the closest distance between the empirical distribution of the test samples and the uncertainty set \mathcal{P}_0 . Our test statistic does not depend on \mathcal{P}_1 , but later we will show that it is asymptotically optimal under the Neyman-Pearson setting, i.e., solves the problem in (9).

From Proposition 3, we have that the test statistic of our robust kernel test can be solved analytically in closed-form with a computational complexity of $\mathcal{O}(m^2 + n^2)$. We then show that the kernel robust test in (38) is asymptotically optimal for the problem in (9) in the following theorem, i.e., it achieves the universal upper bound on the worst-case error exponent in Proposition 5.

Theorem 4. *The robust kernel test in (38) is asymptotically optimal under Neyman-Pearson setting:*

1) under H_0 ,

$$\sup_{P_0 \in \mathcal{P}_0} E_{P_0}[\phi_N(x^n)] = \sup_{P_0 \in \mathcal{P}_0} P_0 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} > \gamma_n \right) \leq \alpha; \quad (39)$$

and 2) under H_1 ,

$$\begin{aligned} & \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi_N(x^n)] \\ &= \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_1 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \leq \gamma_n \right) \\ &= \sup_{\phi: P_F(\phi) \leq \alpha} \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi(x^n)] \\ &= \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1). \end{aligned} \quad (40)$$

Proof. The proof of Theorem 4 can be found in Appendix H. \square

The optimality result for the kernel robust test (38) in Theorem 4 holds for general robust hypothesis testing problems, i.e., it applies to robust hypothesis testing problems defined using different uncertainty sets $\mathcal{P}_0, \mathcal{P}_1$ and using any arbitrary nominal distributions. However, to solve the optimization problem in the test statistic $\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}}$ for any arbitrary uncertainty set \mathcal{P}_0 , it may not always be tractable, since it is an infinite-dimensional problem.

V. SIMULATION RESULTS

In this section, we provide some numerical results to demonstrate the performance of our proposed tests.

A. A Toy Example

We first provide a toy example to visualize the impact of our kernel robust framework. Assume the whole space $\mathcal{X} = \{1, 2, 3, 4, 5\}$. Under hypothesis H_0 , the training samples are 1, 2, 3. Under hypothesis H_1 , the training samples are 3, 4, 5. The radius is set to be 0.2. We choose a Gaussian kernel. The bandwidth for the Gaussian kernel is chosen using the medium heuristic [26]. We plot the empirical distributions and the least favorable distributions (LFDs). The supports of the empirical distributions overlap only at $x = 3$. Comparing the empirical distributions and the LFDs, it can be seen from Fig. 1 that under H_0 , part of the probability mass is transported from $\{1, 2\}$ to $\{4, 5\}$, under H_1 , part of the probability mass is transported from $\{4, 5\}$ to $\{1, 2\}$. Therefore, the LFDs are more difficult to distinguish than the empirical distributions.

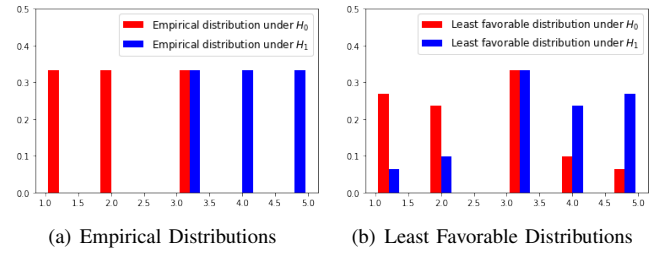


Fig. 1. Comparison of the Empirical Distributions and the Least Favorable Distributions on a Toy Example.

B. Exponential Consistency of the Tests

In this section, we validate the exponential consistency of the proposed tests. We use 40 samples from $\mathcal{N}(0, \mathbf{I})$ and 40 samples from $\mathcal{N}(0.22\mathbf{e}, \mathbf{I})$ to construct the uncertainty sets under H_0 and H_1 respectively, where \mathbf{e} is a vector with all entries equal to 1, and \mathbf{I} is the identity matrix. The data dimension is 20. The radii are chosen such that the uncertainty sets do not overlap. For the kernel smoothing robust test, we use training samples as the support of the finite-dimensional robust optimization problem in (17). When testing the batch samples, we take the sum of the log-likelihood ratio for each sample and compare it with a threshold. We choose the Gaussian kernel and the bandwidth parameter is chosen using cross-validation. We use the data-generating distributions to evaluate the performance of the two tests. We plot the log of the error probability under the Bayesian setting as a function of testing sample size n . It can be seen from Fig. 1 that with the increasing of sample size n , the error probabilities of the direct robust test and kernel smoothing robust test decay exponentially fast, which validates the theoretical result that the direct robust test is exponentially consistent. Moreover, our kernel smoothing robust test has a better performance than the direct robust kernel test.

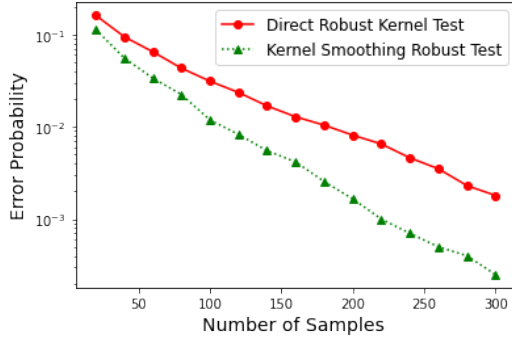


Fig. 2. Exponential Consistency of the Tests.

C. Comparison of the Performance

In this section, we compare our tests with the Wasserstein robust test [9], [10].

We first compare the performance using synthetic data. We use 20 samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and 20 samples from $\mathcal{N}(\mathbf{e}, \mathbf{I})$ to construct the uncertainty sets under H_0 and H_1 respectively. The data dimension is 4. We use a Gaussian kernel and the bandwidth parameter is chosen using the cross-validation. For a fair comparison, and due to the difficulty of obtaining the coefficients in the Wasserstein distance concentration bound in [10, Section 4], we compute the distance between the true distribution and the empirical distribution of the training samples using Monte Carlo method and use it as the radii of the uncertainty sets so that the true distributions lie in the uncertainty sets. We then use the true distributions to evaluate the performance of the proposed tests. We plot the log of the error probability as a function of testing sample size n . It can be seen from Fig. 3 that the kernel smoothing robust test has the best performance. The performance of the direct robust test and the Wasserstein robust test are close.

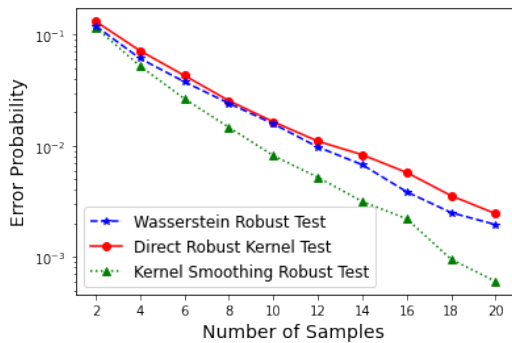


Fig. 3. Comparison of the Kernel Smoothing Robust Test, the Direct Robust Kernel Test and the Wasserstein Robust Test: Synthetic Data.

We then validate the performance of our robust tests using real data of human activity detection. The dataset was released by the Wireless Sensor Data Mining (WISDM) Lab in October 2013, which was collected with the Actitracker system [55]–[57]. Users carried smartphone and were asked to do different activities. For each person, the dataset records the user's name, activities and the acceleration of the user in three directions.

We use the walking data and the jogging data collected from four different users to form H_0 and H_1 respectively. We use five samples from each user to construct the uncertainty sets. The radii of the uncertainty sets are chosen by cross-validation for fair comparison. We plot the log scale error probability as a function of testing sample size n . In Fig. 4, it can be seen that the performance of the kernel smoothing robust test is better than the Wasserstein robust test and the direct robust kernel test. These results demonstrate the good performance of our kernel robust framework.

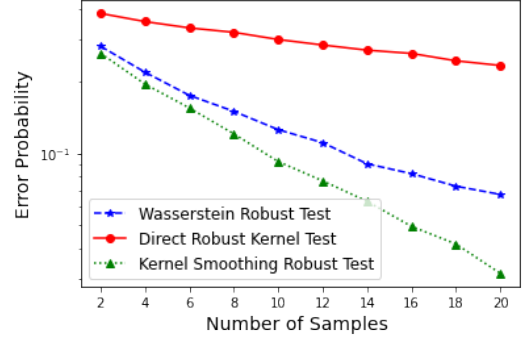


Fig. 4. Comparison of the Kernel Smoothing Robust Test, the Direct Robust Kernel Test and the Wasserstein Robust Test: Human Activity Dataset.

We then compare the performance of the three algorithms using MNIST handwritten digits dataset [58]. We first normalize the image data and then select five images from two different classes to construct the uncertainty sets. The radii of the uncertainty sets are chosen by cross-validation for fair comparison. We plot the log scale error probability as a function of testing sample size n . From Fig. 5, it can be seen that the performance of the kernel smoothing robust test and the direct robust kernel test are close. Moreover, both the kernel smoothing robust test and the direct robust kernel test outperform the Wasserstein robust test.

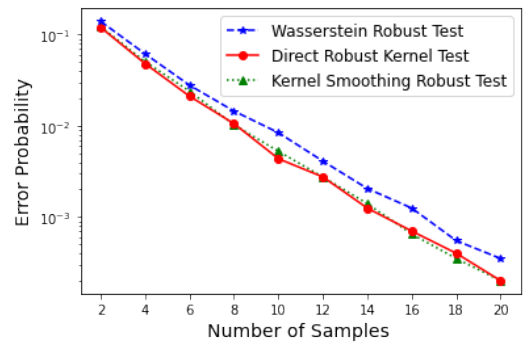


Fig. 5. Comparison of the Kernel Smoothing Robust Test, the Direct Robust Kernel Test and the Wasserstein Robust Test: MNIST Handwritten Digits Dataset.

D. Comparison under Different Dimensions and Training Sample Sizes

In this section, we compare the performance of our kernel smoothing robust test under different data dimensions and

different training sample sizes.

We evaluate the performance of our kernel smoothing test when the data dimension is 5, 10, 15 and 20. When the data dimension is 5, we use 20 samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and 20 samples from $\mathcal{N}(0.48\mathbf{e}, \mathbf{I})$ to construct the uncertainty sets under H_0 and H_1 respectively. When the data dimensions are 10, 15 and 20, we scale the mean of the Gaussian distribution under H_1 so that the KL divergence between the true distributions under H_0 and H_1 is the same for different data dimensions [59]. We use a Gaussian kernel and the bandwidth parameter is chosen using cross-validation. It can be seen from Fig. 6 that the performance of the kernel smoothing test decreases when the data dimension increases.

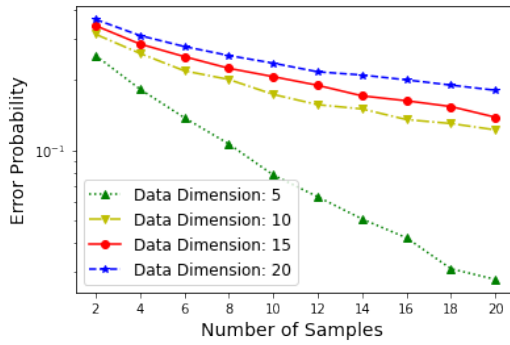


Fig. 6. Comparison of the Kernel Smoothing Robust Test under Different Data Dimensions.

We then examine the impact of the training sample size on our kernel smoothing test. We use different number of training samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(0.22\mathbf{e}, \mathbf{I})$ to construct the uncertainty sets under H_0 and H_1 respectively. The data dimension is 20. The radius of the uncertainty set is chosen such that the true distributions lie in the uncertainty sets with the same probability for different training sample sizes. From Fig. 7, it can be seen that the kernel smoothing test performs better when the training sample size is larger. This validates the observation that when the training sample size is larger, we have more information about the true distributions, and the problem shall be easier to solve.

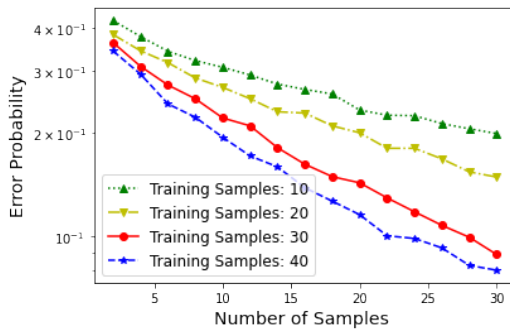


Fig. 7. Comparison of the Kernel Smoothing Robust Test under Different Training Sample Sizes.

E. Robust Kernel Test under the Neyman-Pearson Setting

For the Neyman-Pearson setting, we show the good performance of our robust kernel test. We first demonstrate the performance of our tests using multivariate Gaussian distributions. For hypotheses H_0 , we use 50 samples generated from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to construct the uncertainty set. For H_1 , we use 50 samples generated from $\mathcal{N}(0.5\mathbf{e}, \mathbf{I})$ to construct the uncertainty set. The data dimension is 4. The radii are chosen such that the uncertainty sets do not overlap. To test the robustness of our tests, we choose an arbitrary pair of distributions P_0, P_1 that lie on the boundary of the uncertainty sets $\mathcal{P}_0, \mathcal{P}_1$. Specifically, P_0 and P_1 are multi-variate Gaussian distributions with mean $0.1\mathbf{e}$ and $0.4\mathbf{e}$, respectively, and with the same covariance matrix.

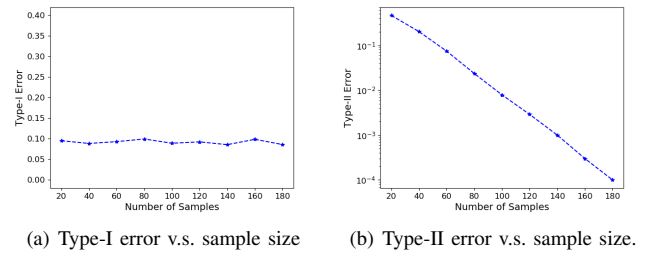


Fig. 8. Error Probability of Robust Kernel Test on Synthetic Dataset.

We set the false alarm constraint $\alpha = 0.1$. With a proper choice of threshold, in Fig. 8(a), we plot the type-I error probability as a function of sample size n . We repeat the experiment for 10000 times. In Fig. 8(b), we plot the the type-II error probability as a function of sample size n . It can be seen that the type-II error probability of our robust kernel test decays exponentially fast with the sample size n while the type-I error probability satisfies the false alarm constraint.

We then use the real data set as in Section V-C to demonstrate the performance of our robust kernel test. We use the walking data collected from the person indexed by 685 and the walking data collected from the person indexed by 669 to form hypotheses H_0 and H_1 . A small portion of the data is used to construct the uncertainty sets. The radius θ of the uncertainty sets is chosen such that the two uncertainty sets do not overlap.

We set the false alarm constraint $\alpha = 0.1$. With a proper choice of threshold, we plot the type-I and type-II error probability as a function of sample size n . From Fig. 9, it can be seen that the type-II error probability of our robust kernel test decays exponentially fast with sample size n while the type-I error probability satisfies the false alarm constraint.

VI. CONCLUSION

In this paper, we studied the robust hypothesis testing problem. We proposed a data-driven approach to construct the uncertainty sets using distance between kernel mean embeddings of distributions. Under the Bayesian setting, we first found the optimal test for the case with a finite alphabet. For the case with an infinite alphabet, we proposed a tractable approximation to quantify the worst-case error probability,

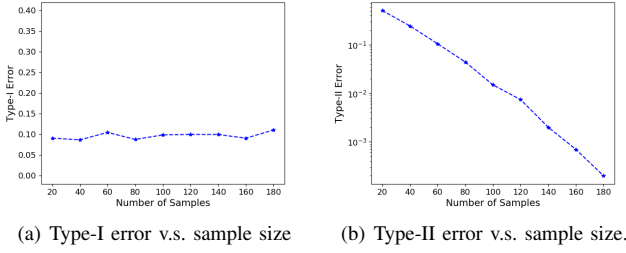


Fig. 9. Error Probability of Robust Kernel Test on the Real Dataset.

and we developed a kernel smoothing method to generalize to unseen data in the alphabet. We also developed a direct robust kernel test which was further shown to be exponentially consistent. Under the Neyman-Pearson setting, we constructed a robust kernel test which can be implemented efficiently and further proved that the proposed test is asymptotically optimal. Specifically, we derived an universal upper bound on the type-II error exponent, and then showed that our robust kernel test achieved this universal upper bound. We also provided some numerical results to demonstrate the performance of our tests. Our approaches provide useful insights for robust hypothesis testing problems in high-dimensional setting.

In the future, it is of interest to investigate the robust multiple hypothesis testing problem with kernel uncertainty sets, where the design of robust detector is significantly more challenging. Another possible extension is to consider the kernel robust sequential hypothesis testing. In this case, we aim to minimize the worst-case probability of errors regarding the hypothesis using as few samples as possible, for which a data-driven approach needs to be developed.

APPENDIX A USEFUL LEMMAS

In this section, we list one useful lemma for our proof.

Lemma 3. [60], [61] For any $P_1 \in \mathcal{P}$, the KL-divergence $D(\cdot \| P_1)$ 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_0 such that for any $P' \in U$, $D(P' \| P_1) \geq D(P_0 \| P_1) - \epsilon$ if $D(P_0 \| P_1) < \infty$, and $D(P' \| P_1) \rightarrow \infty$ as P' converges to P_0 if $D(P_0 \| P_1) = \infty$.

APPENDIX B RADI SELECTION

We first provide a concentration results for kernel MMD in the following lemma.

Lemma 4. [27], [28] Assume $0 \leq k(\cdot, \cdot) \leq K$. Given samples $x^m = (x_1, x_2, \dots, x_m)$ i.i.d. generated from P_0 , denote by \hat{P}_m the empirical distribution of x^m , we then have that

$$P_0\left(\left\|\mu_{\hat{P}_m} - \mu_{P_0}\right\|_{\mathcal{H}} > (2K/m)^{1/2} + \epsilon\right) \leq \exp\left(-\frac{\epsilon^2 m}{2K}\right).$$

This lemma provides a method to choose the radius of the uncertainty set so that the true distribution lies in the uncertainty set with high probability. Let $0 < \delta < 1$, when

the training sample size in m , to guarantee that the true distribution lies in the uncertainty set with probability at least $1 - \delta$, the radius of the uncertainty set should be chosen as

$$\theta = \sqrt{\frac{2K}{m}} + \sqrt{\frac{2K \log \frac{1}{\delta}}{m}}. \quad (41)$$

This method in (41) is a straightforward approach to apply, and usually works very when there is a good number of training samples. In practice, to avoid being overly conservative, it is recommended to choose the radii using this method together with approaches, e.g., cross validation.

APPENDIX C PROOF OF LEMMA 2

Proof. To prove Lemma 2, we will first show that $\int \min\{p_0(x), p_1(x)\}dx$ is concave in p_0, p_1 . Let B be the σ -field on \mathcal{X} . Let $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{|\mathcal{A}|}\}$ be a finite partition of \mathcal{X} which divides \mathcal{X} into a finite number of sets and $|\mathcal{A}|$ denotes the number of partitions in \mathcal{A} . Denote by Π the collection of all finite B-measurable partitions. Let $P_0^{\mathcal{A}_i} = P_0(\mathcal{A}_i)$ and $P_1^{\mathcal{A}_i} = P_1(\mathcal{A}_i)$ for $i = 1, 2, \dots, |\mathcal{A}|$. We will then prove that $\int \min\{p_0(x), p_1(x)\}dx = \inf_{\mathcal{A} \in \Pi} \sum_{i=1}^{|\mathcal{A}|} \min\{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\}$, and show the upper semi-continuity.

Step 1. Let $f_0(p_0, p_1) = p_0$ and $f_1(p_0, p_1) = p_1$. Since $f_0(p_0, p_1)$ and $f_1(p_0, p_1)$ are linear in p_0, p_1 , $\min\{p_0, p_1\}$ is the minimum of two linear functions thus is concave. Therefore, $\int \min\{p_0(x), p_1(x)\}dx$ is concave in p_0, p_1 .

Step 2. For any partitions $\mathcal{A} \in \Pi$, we have that $\mathcal{A}_i \in B, \forall i \in \{1, 2, \dots, |\mathcal{A}|\}$. For any $\mathcal{A} \in \Pi$, from the concavity of $\min\{p_0(x), p_1(x)\}$ and Jensen's inequality [52], we have that

$$\int \min\{p_0(x), p_1(x)\}dx \leq \sum_{i=1}^{|\mathcal{A}|} \min\{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\}. \quad (42)$$

We note that $0 \leq \min\{\frac{p_0(x)}{p_1(x)}, 1\} \leq 1$. Therefore, for any $\epsilon > 0$, there exists a partition $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{|\mathcal{A}|}\}$ such that $\bigcup_{i=1}^{|\mathcal{A}|} \mathcal{A}_i = \mathcal{X}$ and $\bar{h}_i - \underline{h}_i < \epsilon$, where $\bar{h}_i = \sup_{x \in \mathcal{A}_i} \min\{\frac{p_0(x)}{p_1(x)}, 1\}$, $\underline{h}_i = \inf_{x \in \mathcal{A}_i} \min\{\frac{p_0(x)}{p_1(x)}, 1\}$ for $i = 1, 2, \dots, |\mathcal{A}|$. We then have that

$$\begin{aligned} \int_{\mathcal{A}_i} \bar{h}_i p_1(x) dx &\geq \int_{\mathcal{A}_i} \min\left\{\frac{p_0(x)}{p_1(x)}, 1\right\} p_1(x) dx \\ &\geq \int_{\mathcal{A}_i} \underline{h}_i p_1(x) dx. \end{aligned} \quad (43)$$

It follows that

$$P_1(\mathcal{A}_i) \bar{h}_i \geq \int_{\mathcal{A}_i} \min\left\{\frac{p_0(x)}{p_1(x)}, 1\right\} p_1(x) dx \geq P_1(\mathcal{A}_i) \underline{h}_i. \quad (44)$$

Moreover,

$$P_1(\mathcal{A}_i) \bar{h}_i \geq P_1(\mathcal{A}_i) \min\left\{\frac{P_0(\mathcal{A}_i)}{P_1(\mathcal{A}_i)}, 1\right\} \geq P_1(\mathcal{A}_i) \underline{h}_i, \quad (45)$$

where the first inequality is because when $\bar{h}_i = 1$, $\min\{\frac{P_0(\mathcal{A}_i)}{P_1(\mathcal{A}_i)}, 1\} \leq 1$, and when $\bar{h}_i = \sup_{x \in \mathcal{A}_i} \frac{p_0(x)}{p_1(x)}$, $\bar{h}_i =$

$\frac{\bar{h}_i \int_{\mathcal{A}_i} p_1(x) dx}{\int_{\mathcal{A}_i} p_1(x) dx} \geq \frac{\int_{\mathcal{A}_i} p_0(x) dx}{\int_{\mathcal{A}_i} p_1(x) dx} = \frac{P_0(\mathcal{A}_i)}{P_1(\mathcal{A}_i)}$. The second inequality can also be proved similarly.

It then follows from (44) and (45) that

$$\left| P_1(\mathcal{A}_i) \min \left\{ \frac{P_0(\mathcal{A}_i)}{P_1(\mathcal{A}_i)}, 1 \right\} - \int_{\mathcal{A}_i} \min \{p_0(x), p_1(x)\} dx \right| \leq (\bar{h}_i - \underline{h}_i) P_1(\mathcal{A}_i) < \epsilon P_1(\mathcal{A}_i). \quad (46)$$

Therefore,

$$\left| \sum_{i=1}^{|\mathcal{A}|} P_1(\mathcal{A}_i) \min \left\{ \frac{P_0(\mathcal{A}_i)}{P_1(\mathcal{A}_i)}, 1 \right\} - \int \min \{p_0(x), p_1(x)\} dx \right| \leq \epsilon. \quad (47)$$

We then have that

$$\sum_{i=1}^{|\mathcal{A}|} P_1(\mathcal{A}_i) \min \left\{ \frac{P_0(\mathcal{A}_i)}{P_1(\mathcal{A}_i)}, 1 \right\} \leq \int \min \{p_0(x), p_1(x)\} dx + \epsilon. \quad (48)$$

Let $\epsilon \rightarrow 0$, we have that

$$\inf_{\mathcal{A} \in \Pi} \sum_{i=1}^{|\mathcal{A}|} \min \{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\} \leq \int \min \{p_0(x), p_1(x)\} dx. \quad (49)$$

Combining (42) and (49), we have that

$$\int \min \{p_0(x), p_1(x)\} dx = \inf_{\mathcal{A} \in \Pi} \sum_{i=1}^{|\mathcal{A}|} \min \{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\}. \quad (50)$$

Step 3. Let C be the field of Borel sets of \mathcal{X} that are sets of continuity for both P_0 and P_1 . It was shown in [62, Theorem 1] that C generates B in the sense that B is the smallest σ -field containing C .

Let $\mathcal{A}^C = \{\mathcal{A}_1^C, \mathcal{A}_2^C, \dots, \mathcal{A}_{|\mathcal{A}|}^C\}$ be a finite partition of \mathcal{X} such that $\mathcal{A}_i^C \in C$, $\forall i \in \{1, 2, \dots, |\mathcal{A}|\}$. Let Π^C be the collection of all such finite partitions. Since $\Pi^C \subseteq \Pi$, we have that

$$\inf_{\mathcal{A}^C \in \Pi^C} \sum_{i=1}^{|\mathcal{A}^C|} \min \{P_0^{\mathcal{A}_i^C}, P_1^{\mathcal{A}_i^C}\} \geq \inf_{\mathcal{A} \in \Pi} \sum_{i=1}^{|\mathcal{A}|} \min \{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\}. \quad (51)$$

Since C generates B , applying Theorem D of section 13 [63] to the measure $\nu = P_0 + P_1$, for any $\epsilon > 0$ and $\mathcal{A} \in \Pi$, we can find $E'_i \in C$ such that $P_0(\mathcal{A}_i \Delta E'_i) \leq \nu(\mathcal{A}_i \Delta E'_i) \leq \epsilon$ and $P_1(\mathcal{A}_i \Delta E'_i) \leq \nu(\mathcal{A}_i \Delta E'_i) \leq \epsilon$, where Δ denotes the symmetric difference between two sets. Define

$$\begin{aligned} E_1 &= E'_1, \\ E_2 &= E'_2 - E_1, \dots, E_{|\mathcal{A}|-1} \\ &= E'_{|\mathcal{A}|-1} - E_1 \cup E_2 \cup \dots \cup E_{|\mathcal{A}|-2}, \end{aligned} \quad (52)$$

and

$$E_{|\mathcal{A}|} = \mathcal{X} - E_1 \cup E_2 \cup \dots \cup E_{|\mathcal{A}|-1}. \quad (53)$$

We have that $\nu(\mathcal{A}_1 \Delta E_1) = \nu(\mathcal{A}_1 \Delta E'_1) \leq \epsilon$. Since $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{|\mathcal{A}|}$ are disjoint, we have that $\nu(\mathcal{A}_2 \cap E_1) \leq \epsilon$. It then follows that

$$\begin{aligned} \nu(\mathcal{A}_2 \Delta E_2) &= \nu(\mathcal{A}_2 \Delta (E'_2 - E_1)) \\ &\leq \nu(\mathcal{A}_2 \Delta E'_2) + \nu(\mathcal{A}_2 \cap E_1) \leq 2\epsilon. \end{aligned} \quad (54)$$

Similarly, we can show that for any $1 \leq i \leq |\mathcal{A}| - 1$,

$$\nu(\mathcal{A}_i \Delta E_i) = \nu(\mathcal{A}_i \Delta (E'_i - E_1 \cup E_2 \cup \dots \cup E_{i-1})) \leq i\epsilon \quad (55)$$

and

$$\begin{aligned} \nu(\mathcal{A}_{|\mathcal{A}|} \Delta E_{|\mathcal{A}|}) &= \nu((\mathcal{X} - \mathcal{A}_1 \cup \dots \cup \mathcal{A}_{|\mathcal{A}|-1}) \Delta (\mathcal{X} - E_1 \cup \dots \cup E_{|\mathcal{A}|-1})) \\ &= \nu((\mathcal{A}_1 \cup \dots \cup \mathcal{A}_{|\mathcal{A}|-1}) \Delta (E_1 \cup \dots \cup E_{|\mathcal{A}|-1})) \\ &\leq \frac{|\mathcal{A}|(|\mathcal{A}| - 1)}{2} \epsilon. \end{aligned} \quad (56)$$

Therefore, for any $\epsilon > 0$ and $\mathcal{A} \in \Pi$, there exists $E_1, E_2, \dots, E_{|\mathcal{A}|} \in C$ such that

$$\begin{aligned} P_0(E_i) &\leq P_0(\mathcal{A}_i) + \frac{|\mathcal{A}|(|\mathcal{A}| - 1)}{2} \epsilon, \\ P_1(E_i) &\leq P_1(\mathcal{A}_i) + \frac{|\mathcal{A}|(|\mathcal{A}| - 1)}{2} \epsilon, \forall 1 \leq i \leq |\mathcal{A}|. \end{aligned} \quad (57)$$

It then follows that there exists $\mathcal{A}^C \in \Pi^C$ such that

$$\begin{aligned} \sum_{i=1}^{|\mathcal{A}^C|} \min \{P_0^{\mathcal{A}_i^C}, P_1^{\mathcal{A}_i^C}\} &\leq \inf_{\mathcal{A} \in \Pi} \sum_{i=1}^{|\mathcal{A}|} \min \{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\} + \frac{|\mathcal{A}|^2(|\mathcal{A}| - 1)}{2} \epsilon. \end{aligned} \quad (58)$$

Let $\epsilon \rightarrow 0$ and use (51), we then have that

$$\inf_{\mathcal{A}^C \in \Pi^C} \sum_{i=1}^{|\mathcal{A}^C|} \min \{P_0^{\mathcal{A}_i^C}, P_1^{\mathcal{A}_i^C}\} = \inf_{\mathcal{A} \in \Pi} \sum_{i=1}^{|\mathcal{A}|} \min \{P_0^{\mathcal{A}_i}, P_1^{\mathcal{A}_i}\}. \quad (59)$$

Therefore,

$$\int \min \{p_0(x), p_1(x)\} dx = \inf_{\mathcal{A}^C \in \Pi^C} \sum_{i=1}^{|\mathcal{A}^C|} \min \{P_0^{\mathcal{A}_i^C}, P_1^{\mathcal{A}_i^C}\}. \quad (60)$$

Let P_0^s and P_1^s be the sequence of probability distribution such that P_0^s converges weakly to P_0 and P_1^s converges weakly to P_1 as $s \rightarrow \infty$. Let $\mathcal{A}^C = \{\mathcal{A}_1^C, \mathcal{A}_2^C, \dots, \mathcal{A}_{|\mathcal{A}|}^C\} \in \Pi^C$. We then have that $\mathcal{A}_i^C \in C$ for all $i \in \{1, 2, \dots, |\mathcal{A}|\}$. From the Portmanteau theorem [64] and the fact that \mathcal{A}_i^C is a continuity set of P_0 and P_1 , the weak convergence implies $P_0^s(\mathcal{A}_i^C) \rightarrow P_0(\mathcal{A}_i^C)$ and $P_1^s(\mathcal{A}_i^C) \rightarrow P_1(\mathcal{A}_i^C)$. It then follows that for any $\epsilon > 0$, there exists s_0 such that for any $s > s_0$,

$$P_0(\mathcal{A}_i^C) + \frac{\epsilon}{|\mathcal{A}^C|} > P_0^s(\mathcal{A}_i^C), P_1(\mathcal{A}_i^C) + \frac{\epsilon}{|\mathcal{A}^C|} > P_1^s(\mathcal{A}_i^C). \quad (61)$$

Therefore, for any $\epsilon > 0$, there exists $\mathcal{A}^C \in \Pi^C$ such that

$$\begin{aligned} & \int \min \{p_0(x), p_1(x)\} dx + 2\epsilon \\ & \geq \sum_{i=1}^{|\mathcal{A}^C|} \min \{P_0^{\mathcal{A}_i^C}, P_1^{\mathcal{A}_i^C}\} + \epsilon \\ & = \sum_{i=1}^{|\mathcal{A}^C|} \min \left\{ P_0^{\mathcal{A}_i^C} + \frac{\epsilon}{|\mathcal{A}^C|}, P_1^{\mathcal{A}_i^C} + \frac{\epsilon}{|\mathcal{A}^C|} \right\} \\ & \geq \lim_{s \rightarrow \infty} \sum_{i=1}^{|\mathcal{A}^C|} \min \{P_0^s(\mathcal{A}_i^C), P_1^s(\mathcal{A}_i^C)\} \\ & \geq \lim_{s \rightarrow \infty} \int \min \{p_0^s(x), p_1^s(x)\} dx, \end{aligned} \quad (62)$$

where the first inequality is from (60), the last inequality is from (42). Let $\epsilon \rightarrow 0$, we have that $\int \min \{p_0(x), p_1(x)\} dx$ is upper semi-continuous in P_0, P_1 with respect to the weak convergence. This completes the proof. \square

APPENDIX D PROOF OF THEOREM 1

Proof. Since \mathcal{X} is compact, \mathcal{P} is tight, thus is sequentially compact with respect to the topology of weak convergence from the Prokhorov's theorem [65]. Therefore, \mathcal{P} is compact with respect to weak convergence. Therefore, $\mathcal{P}_0, \mathcal{P}_1$ are compact since $\mathcal{P}_0, \mathcal{P}_1$ are closed subsets of a compact set. We then have that the solutions to $\sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \int \min \{p_0(x), p_1(x)\} dx$ exist because upper semi-continuous function attains its supremum on a compact set. Let P_0^*, P_1^* denote the optimal solutions to $\sup_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} \int \min \{p_0(x), p_1(x)\} dx$.

Let $\mathcal{A}_s = \{\mathcal{A}_s^1, \mathcal{A}_s^2, \dots, \mathcal{A}_s^s\}$ be a partition of \mathcal{X} . We define the diameter of each partition \mathcal{A}_s^i as $\text{dia}(\mathcal{A}_s^i) = \max_{x, x' \in \mathcal{A}_s^i} \|x - x'\|_2$. Since \mathcal{X} is compact, we can choose the partition \mathcal{A}_s such that $\text{dia}(\mathcal{A}_s^i) \rightarrow 0$ as $s \rightarrow \infty$ for $1 \leq i \leq s$.

For any partition \mathcal{A}_s^j , let x_s^j be an arbitrary point in \mathcal{A}_s^j . Denote by P_0^s, P_1^s discrete distributions with $P_0^s(x_s^j) = P_0^*(\mathcal{A}_s^j), P_1^s(x_s^j) = P_1^*(\mathcal{A}_s^j)$. Let $h : \mathcal{X} \rightarrow \mathbb{R}$ be an arbitrary bounded, continuous function. Let $a_s^j = \inf_{x \in \mathcal{A}_s^j} h(x)$ and $b_s^j = \sup_{x \in \mathcal{A}_s^j} h(x)$. Since h is continuous and the diameter of \mathcal{A}_s^j goes to 0 as $s \rightarrow \infty$, we then have that $\max_{j=1, \dots, s} (b_s^j - a_s^j) \rightarrow 0$ as $s \rightarrow \infty$. It then follows that

$$\begin{aligned} & \left| \int h dP_0^s - \int h dP_0^* \right| = \left| \sum_{j=1}^s \int_{\mathcal{A}_s^j} (h - h(x_s^j)) dP_0^* \right| \\ & \leq \max_{j=1, \dots, s} (b_s^j - a_s^j) \rightarrow 0, \text{ as } s \rightarrow \infty. \end{aligned} \quad (63)$$

Therefore, P_0^s converges weakly to P_0^* as $s \rightarrow \infty$. Similarly, P_1^s converges weakly to P_1^* as $s \rightarrow \infty$. Moreover, from Jensen's inequality [52], we have that

$$\begin{aligned} & \int \min \{p_0^*(x), p_1^*(x)\} dx \leq \sum_{j=1}^s \min \{P_0^*(\mathcal{A}_s^j), P_1^*(\mathcal{A}_s^j)\} \\ & = \sum_{j=1}^s \min \{P_0^s(x_s^j), P_1^s(x_s^j)\}. \end{aligned} \quad (64)$$

Since MMD metrizes the weak convergence [45], [46], for any $\epsilon > 0$, there exists an integer s_0 such that for any $s > s_0$, $\|\mu_{P_0^s} - \mu_{P_0^*}\|_{\mathcal{H}} \leq \frac{\epsilon}{2}$ and $\|\mu_{P_1^s} - \mu_{P_1^*}\|_{\mathcal{H}} \leq \frac{\epsilon}{2}$. Therefore, from the triangle inequality [26], we have that for any $s > s_0$,

$$\begin{aligned} \|\mu_{P_0^s} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} & \leq \|\mu_{P_0^s} - \mu_{P_0^*}\|_{\mathcal{H}} + \|\mu_{P_0^*} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \leq \theta + \frac{\epsilon}{2}. \end{aligned} \quad (65)$$

Similarly, we have $\|\mu_{P_1^s} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \frac{\epsilon}{2}$.

Rewrite $P_0^s = \sum_{j=1}^s \alpha_j \delta_{x_s^j}, P_1^s = \sum_{j=1}^s \beta_j \delta_{x_s^j}$, where $\delta_{x_s^j}$ denote the Dirac measure on x_s^j , $\alpha_j = P_0^*(\mathcal{A}_s^j), \beta_j = P_1^*(\mathcal{A}_s^j), \forall j = 1, \dots, s$, and $\sum_{j=1}^s \alpha_j = 1, \sum_{j=1}^s \beta_j = 1$. Let $\text{dis}(x, y)$ denote a distance metric on \mathcal{X} between x and y . Note that $\{z_i\}_{i=1}^N$ are generated from a distribution P supported on \mathcal{X} . Therefore, for any x_s^j , we have that $\min_{z \in \{z_i\}_{i=1}^N} \text{dis}(z, x_s^j) \rightarrow 0$ as $N \rightarrow \infty$. Therefore, there exists a sequence $\{z_{sj}^N\}_{N=1}^\infty$ such that $z_{sj}^N \in \{z_i\}_{i=1}^N$ and $\text{dis}(z_{sj}^N, x_s^j) \rightarrow 0$ as $N \rightarrow \infty$. Assume that z_{sj}^N are distinct for all j . We then construct the following distributions: $P_0^{s,N} = \sum_{j=1}^s \alpha_j \delta_{z_{sj}^N}, P_1^{s,N} = \sum_{j=1}^s \beta_j \delta_{z_{sj}^N}$. For any arbitrary bounded, continuous function $h : \mathcal{X} \rightarrow \mathbb{R}$, since $\text{dis}(z_{sj}^N, x_s^j) \rightarrow 0$ as $N \rightarrow \infty$, we have that for a fixed s ,

$$\begin{aligned} & \left| \int h dP_0^{s,N} - \int h dP_0^s \right| = \left| \sum_{j=1}^s \alpha_j (h(z_{sj}^N) - h(x_s^j)) \right| \\ & \leq \max_{j=1, \dots, s} |h(z_{sj}^N) - h(x_s^j)| \rightarrow 0, \text{ as } N \rightarrow \infty. \end{aligned} \quad (66)$$

Therefore, we have that $P_0^{s,N}$ converges weakly to P_0^s as $N \rightarrow \infty$. Similarly, $P_1^{s,N}$ converges weakly to P_1^s as $N \rightarrow \infty$. Moreover, we have that

$$\begin{aligned} & \sum_{j=1}^s \min \{P_0^{s,N}(z_{sj}^N), P_1^{s,N}(z_{sj}^N)\} \\ & = \sum_{j=1}^s \min \{P_0^s(x_s^j), P_1^s(x_s^j)\} \\ & = \sum_{j=1}^s \min \{\alpha_j, \beta_j\}. \end{aligned} \quad (67)$$

Since P_0^s, P_1^s converges weakly to P_0^*, P_1^* , respectively, as $s \rightarrow \infty$, and $P_0^{s,N}, P_1^{s,N}$ converges weakly to P_0^s, P_1^s , respectively, as $N \rightarrow \infty$, we have that for any $\epsilon > 0$, there exists an integer s_0 such that for all $s > s_0$, $\|\mu_{P_0^s} - \mu_{P_0^*}\|_{\mathcal{H}} \leq \frac{\epsilon}{2}$ and $\|\mu_{P_1^s} - \mu_{P_1^*}\|_{\mathcal{H}} \leq \frac{\epsilon}{2}$. For a fixed $s > s_0$, there exists an integer $N(s)$ such that for any $N > N(s)$, $\|\mu_{P_0^{s,N}} - \mu_{P_0^s}\|_{\mathcal{H}} \leq \frac{\epsilon}{2}$ and $\|\mu_{P_1^{s,N}} - \mu_{P_1^s}\|_{\mathcal{H}} \leq \frac{\epsilon}{2}$. Therefore, for a fixed $s > s_0$ and any $N > N(s)$, from the triangle inequality [26], we have that

$$\begin{aligned} & \|\mu_{P_0^{s,N}} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \leq \|\mu_{P_0^{s,N}} - \mu_{P_0^s}\|_{\mathcal{H}} + \|\mu_{P_0^s} - \mu_{P_0^*}\|_{\mathcal{H}} + \|\mu_{P_0^*} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \leq \theta + \epsilon. \end{aligned} \quad (68)$$

Similarly, we have that $\|\mu_{P_1^s, N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon$. It then follows that for large N ,

$$\begin{aligned} & \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\} \\ & \geq \sum_{j=1}^s \min \{P_0^{s, N}(z_{sj}^N), P_1^{s, N}(z_{sj}^N)\} \\ & \geq \int \min \{p_0^*(x), p_1^*(x)\} dx, \end{aligned} \quad (69)$$

where the second inequality is from Jensen's inequality [52]. Therefore, for any $\epsilon > 0$,

$$\begin{aligned} & \lim_{N \rightarrow \infty} \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\} \\ & \geq \int \min \{p_0^*(x), p_1^*(x)\} dx. \end{aligned} \quad (70)$$

Moreover, we have that for any $\epsilon > 0$,

$$\begin{aligned} & \sup_{\substack{P_0 \in \mathcal{P}: \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1 \in \mathcal{P}: \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon}} \int \min \{p_0(x), p_1(x)\} dx \\ & \geq \lim_{N \rightarrow \infty} \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\}, \end{aligned} \quad (71)$$

which is due to the fact that the right-hand side and the left-hand side of (71) have the same objective function and the feasible region of the right-hand side is a subset of the feasible region of the left-hand side. It then follows that

$$\begin{aligned} & \lim_{\epsilon \rightarrow 0} \sup_{\substack{P_0 \in \mathcal{P}: \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1 \in \mathcal{P}: \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon}} \int \min \{p_0(x), p_1(x)\} dx \\ & \geq \lim_{\epsilon \rightarrow 0} \lim_{N \rightarrow \infty} \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\} \\ & \geq \int \min \{p_0^*(x), p_1^*(x)\} dx. \end{aligned} \quad (72)$$

We will then show that all the inequality holds with equality in (72). Recall the definition of $g(\theta)$ in (19):

$$g(\theta) = \sup_{\substack{P_0 \in \mathcal{P}: \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta \\ P_1 \in \mathcal{P}: \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta}} \int \min \{p_0(x), p_1(x)\} dx. \quad (73)$$

We will show that $\lim_{\epsilon \rightarrow 0} g(\theta + \epsilon) = g(\theta)$, thus

$$\begin{aligned} & \lim_{\epsilon \rightarrow 0} \sup_{\substack{P_0 \in \mathcal{P}: \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1 \in \mathcal{P}: \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon}} \int \min \{p_0(x), p_1(x)\} dx \\ & = \int \min \{p_0^*(x), p_1^*(x)\} dx. \end{aligned} \quad (74)$$

It suffices to show that $g(\theta)$ is continuous in θ . To show that, we will show that $g(\theta)$ is concave in θ . Let $P_{0, \theta_1}, P_{1, \theta_1}$ be the optimal solutions to $g(\theta_1)$ and $P_{0, \theta_2}, P_{1, \theta_2}$ be the optimal solutions to $g(\theta_2)$. Consider $\lambda P_{0, \theta_1} + (1 - \lambda)P_{0, \theta_2}, \lambda P_{1, \theta_1} + (1 - \lambda)P_{1, \theta_2}$ for $0 < \lambda < 1$. From the triangle inequality [26], we have that

$$\begin{aligned} & \|\lambda \mu_{P_{0, \theta_1}} + (1 - \lambda) \mu_{P_{0, \theta_2}} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \leq \lambda \|\mu_{P_{0, \theta_1}} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} + (1 - \lambda) \|\mu_{P_{0, \theta_2}} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \leq \lambda \theta_1 + (1 - \lambda) \theta_2. \end{aligned} \quad (75)$$

Similarly, we have that $\|\lambda \mu_{P_{1, \theta_1}} + (1 - \lambda) \mu_{P_{1, \theta_2}} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \lambda \theta_1 + (1 - \lambda) \theta_2$. Therefore, $\lambda P_{0, \theta_1} + (1 - \lambda)P_{0, \theta_2}, \lambda P_{1, \theta_1} + (1 - \lambda)P_{1, \theta_2}$ are feasible solutions to $g(\lambda \theta_1 + (1 - \lambda) \theta_2)$. It then follows that

$$\begin{aligned} & g(\lambda \theta_1 + (1 - \lambda) \theta_2) \\ & \geq \int \min \{\lambda p_{0, \theta_1}(x) + (1 - \lambda) p_{0, \theta_2}(x), \\ & \quad \lambda p_{1, \theta_1}(x) + (1 - \lambda) p_{1, \theta_2}(x)\} dx \\ & \geq \lambda \int \min \{p_{0, \theta_1}(x), p_{1, \theta_1}(x)\} dx \\ & \quad + (1 - \lambda) \int \min \{p_{0, \theta_2}(x), p_{1, \theta_2}(x)\} dx \\ & = \lambda g(\theta_1) + (1 - \lambda) g(\theta_2). \end{aligned} \quad (76)$$

Therefore, $g(\theta)$ is concave in θ , and thus is continuous in θ . From (72) and the continuity of $g(\theta)$, we have that for any $\theta \geq 0$,

$$\begin{aligned} & \lim_{\epsilon \rightarrow 0} \sup_{\substack{P_0 \in \mathcal{P}: \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1 \in \mathcal{P}: \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon}} \int \min \{p_0(x), p_1(x)\} dx \\ & = \lim_{\epsilon \rightarrow 0} \lim_{N \rightarrow \infty} \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\} \\ & = \int \min \{p_0^*(x), p_1^*(x)\} dx. \end{aligned} \quad (77)$$

We will then show that

$$\begin{aligned}
 & \lim_{N \rightarrow \infty} \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\} \\
 &= \lim_{\epsilon \rightarrow 0} \lim_{N \rightarrow \infty} \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta + \epsilon \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\}. \quad (78)
 \end{aligned}$$

Recall the definition of $g_N(\theta)$ in (20):

$$\begin{aligned}
 & g_N(\theta) \\
 &= \sup_{\substack{P_0^N \in \mathcal{P}: \|\mu_{P_0^N} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta \\ P_1^N \in \mathcal{P}: \|\mu_{P_1^N} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta \\ P_0^N, P_1^N \text{ are supported on } \{z_i\}_{i=1}^N}} \sum_{i=1}^N \min \{P_0^N(z_i), P_1^N(z_i)\}.
 \end{aligned}$$

Let $g^*(\theta) = \lim_{N \rightarrow \infty} g_N(\theta)$. This limit exists because for any $\theta > 0$, $\{g_N(\theta)\}_{N=1}^{\infty}$ is a non-decreasing sequence and has upper bound $g(\theta)$.

For any N , denote by $P_{0,\theta_1}^N, P_{1,\theta_1}^N$ the optimal solutions to $g_N(\theta_1)$ and $P_{0,\theta_2}^N, P_{1,\theta_2}^N$ the optimal solutions to $g_N(\theta_2)$. Consider $\lambda P_{0,\theta_1}^N + (1-\lambda)P_{0,\theta_2}^N, \lambda P_{1,\theta_1}^N + (1-\lambda)P_{1,\theta_2}^N$ for $0 < \lambda < 1$. We have that

$$\begin{aligned}
 & g^*(\lambda\theta_1 + (1-\lambda)\theta_2) \\
 &= \lim_{N \rightarrow \infty} g_N(\lambda\theta_1 + (1-\lambda)\theta_2) \\
 &\geq \lim_{N \rightarrow \infty} \sum_{i=1}^N \min \{\lambda P_{0,\theta_1}^N(z_i) + (1-\lambda)P_{0,\theta_2}^N(z_i), \\
 &\quad \lambda P_{1,\theta_1}^N(z_i) + (1-\lambda)P_{1,\theta_2}^N(z_i)\} \\
 &\geq \lim_{N \rightarrow \infty} \lambda \sum_{i=1}^N \min \{P_{0,\theta_1}^N(z_i), P_{1,\theta_1}^N(z_i)\} \\
 &\quad + (1-\lambda) \sum_{i=1}^N \min \{P_{0,\theta_2}^N(z_i), P_{1,\theta_2}^N(z_i)\} \\
 &= \lambda \lim_{N \rightarrow \infty} \sum_{i=1}^N \min \{P_{0,\theta_1}^N(z_i), P_{1,\theta_1}^N(z_i)\} \\
 &\quad + (1-\lambda) \lim_{N \rightarrow \infty} \sum_{i=1}^N \min \{P_{0,\theta_2}^N(z_i), P_{1,\theta_2}^N(z_i)\} \\
 &= \lambda g^*(\theta_1) + (1-\lambda)g^*(\theta_2), \quad (79)
 \end{aligned}$$

where the first equality is because the limits $\lim_{N \rightarrow \infty} \sum_{i=1}^N \min \{P_{0,\theta_1}^N(z_i), P_{1,\theta_1}^N(z_i)\}$ and $\lim_{N \rightarrow \infty} \sum_{i=1}^N \min \{P_{0,\theta_2}^N(z_i), P_{1,\theta_2}^N(z_i)\}$ exist. Therefore, $g^*(\theta)$ is concave in θ , and thus is continuous in θ . From the continuity of $g^*(\theta)$ and (77), we have that (78) holds. This completes the proof. \square

APPENDIX E PROOF OF PROPOSITION 3

Proof. Note that MMD is non-negative. If $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta$, we have that $\hat{P}_n \in \mathcal{P}_0$. Therefore, $\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} = 0$. We then consider the case that $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} > \theta$. For any $P \in \mathcal{P}_0$, $\|\mu_P - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta$, and thus by the triangle inequality [26], we have that

$$\begin{aligned}
 \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} &\geq \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_P - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\
 &\geq \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta. \quad (80)
 \end{aligned}$$

It then follows that

$$\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \geq \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta. \quad (81)$$

The equality in (81) can be achieved when the following condition holds for a $P \in \mathcal{P}_0$:

$$\|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} = \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta. \quad (82)$$

We then construct such a P . Let $\lambda = \frac{\theta}{\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}}$. Since $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} > \theta$, we have that $0 < \lambda < 1$. Let $P = \lambda \hat{P}_n + (1-\lambda)\hat{Q}_m^0$ be a linear combination of two distributions \hat{P}_n and \hat{Q}_m^0 . We then have that $P \in \mathcal{P}$ and

$$\begin{aligned}
 \mu_P &= \int k(x, \cdot) d(\lambda \hat{P}_n + (1-\lambda)\hat{Q}_m^0) \\
 &= \lambda \int k(x, \cdot) d\hat{P}_n + (1-\lambda) \int k(x, \cdot) d\hat{Q}_m^0 \\
 &= \lambda \mu_{\hat{P}_n} + (1-\lambda)\mu_{\hat{Q}_m^0}. \quad (83)
 \end{aligned}$$

It then follows that

$$\begin{aligned}
 \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} &= \|(1-\lambda)\mu_{\hat{P}_n} - (1-\lambda)\mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\
 &= (1-\lambda)\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\
 &= \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta. \quad (84)
 \end{aligned}$$

and

$$\begin{aligned}
 \|\mu_P - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} &= \|\lambda \mu_{\hat{P}_n} - \lambda \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\
 &= \lambda \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\
 &= \theta. \quad (85)
 \end{aligned}$$

Therefore, $P = \lambda \hat{P}_n + (1-\lambda)\hat{Q}_m^0 \in \mathcal{P}_0$ and achieves the equality in (81). Therefore, when $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} > \theta$, we have that

$$\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} = \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta, \quad (86)$$

From (2), it follows that

$$\begin{aligned}
 \inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} &= \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta \\
 &= \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(\hat{x}_{0,i}, \hat{x}_{0,j}) \right. \\
 &\quad \left. - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, \hat{x}_{0,j}) \right)^{1/2} - \theta. \quad (87)
 \end{aligned}$$

Following the same idea as in solving $\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}}$, the closed-form solution can also be derived for $\inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}}$. \square

APPENDIX F PROOF OF THEOREM 3

Proof. 1) For the type-I error exponent, from the Sanov's theorem [4], we have that for any $P_0 \in \mathcal{P}_0$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} -\frac{1}{n} E_{P_0}[\phi_B(x^n)] \\ &= \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_0 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \right. \\ & \quad \left. - \inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \geq \gamma \right) \\ &\geq \inf_{P' \in \Gamma_0} D(P' \| P_0), \end{aligned} \quad (88)$$

where $\Gamma_0 = \left\{ P' : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \geq \gamma \right\}$ and $D(P' \| P_0)$ denotes the KL-divergence between two distributions P' and P_0 . For any $P_0 \in \mathcal{P}_0$ and $\gamma \in \left(-\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + 2\theta, \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \right)$, we have that

$$\begin{aligned} & \inf_{P \in \mathcal{P}_0} \|\mu_{P_0} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{P_0} - \mu_P\|_{\mathcal{H}} \\ &= - \inf_{P \in \mathcal{P}_1} \|\mu_{P_0} - \mu_P\|_{\mathcal{H}} \\ &= -\|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + \theta \\ &\leq \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + \theta \\ &\leq -\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + 2\theta \\ &< \gamma, \end{aligned} \quad (89)$$

where the first and second equalities are from Proposition 3, the first inequality is from the triangle inequality of MMD [26] and the second inequality is because $P_0 \in \mathcal{P}_0$. We then have that for any $P_0 \in \mathcal{P}_0$, when $\gamma \in \left(-\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + 2\theta, \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \right)$, $P_0 \notin \Gamma_0$. Therefore, $\inf_{P' \in \Gamma_0} D(P' \| P_0) > 0$ thus the type-I error probability of ϕ_B decreases exponentially fast with n .

Similarly, for the type-II error exponent, we have that for any $P_1 \in \mathcal{P}_1$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1}[1 - \phi_B(x^n)] \\ &= \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_1 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \right. \\ & \quad \left. - \inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} < \gamma \right) \\ &\geq \inf_{P' \in \Gamma_1} D(P' \| P_1), \end{aligned} \quad (90)$$

where $\Gamma_1 = \left\{ P' : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma \right\}$. For any $P_1 \in \mathcal{P}_1$ and $\gamma \in \left(-\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + \right.$

$2\theta, \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \big)$, we have that

$$\begin{aligned} & \inf_{P \in \mathcal{P}_0} \|\mu_{P_1} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{P_1} - \mu_P\|_{\mathcal{H}} \\ &= \inf_{P \in \mathcal{P}_0} \|\mu_{P_1} - \mu_P\|_{\mathcal{H}} \\ &= \|\mu_{P_1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta \\ &\geq \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - \|\mu_{P_1} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - \theta \\ &\geq \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \\ &> \gamma, \end{aligned} \quad (91)$$

where the first and second equalities are from Proposition 3, the first inequality is from the triangle inequality of MMD [26] and the second inequality is because $P_1 \in \mathcal{P}_1$. Therefore, for any $P_1 \in \mathcal{P}_1$, we have that $P_1 \notin \Gamma_1$, and thus $\inf_{P' \in \Gamma_1} D(P' \| P_1) > 0$ and the type-II error probability of ϕ_B decreases exponentially fast with n . Therefore, the direct robust kernel test ϕ_B is exponentially consistent.

2) We will then prove that with $\gamma \in \left(-\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + 2\theta, \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \right)$, ϕ_B and ϕ'_B are equivalent. When $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta$, we have that $\hat{P}_n \in \mathcal{P}_0$. From (89), it follows that

$$\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} < \gamma. \quad (92)$$

Moreover, from triangle inequality [26], we have that

$$\begin{aligned} & \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \\ &\leq \theta - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \\ &\leq \theta - \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ &\leq -\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + 2\theta \\ &< \gamma. \end{aligned} \quad (93)$$

Therefore, when $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta$, $\phi_B = \phi'_B = 0$.

When $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta$, we have that $\hat{P}_n \in \mathcal{P}_1$. From (91), it follows that

$$\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} > \gamma. \quad (94)$$

From the triangle inequality [26], we have that

$$\begin{aligned} & \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \\ &\geq \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \theta \\ &\geq \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - \theta \\ &\geq \|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - 2\theta \\ &> \gamma. \end{aligned} \quad (95)$$

Therefore, when $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} \leq \theta$, $\phi_B = \phi'_B = 1$.

When $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} > \theta$ and $\|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} > \theta$, from Proposition 3, we have that

$$\begin{aligned} & \inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} - \inf_{P \in \mathcal{P}_1} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \\ &= \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - \|\mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}}. \end{aligned} \quad (96)$$

Combining the three different cases, we have that ϕ_B and ϕ'_B are equivalent. We note that ϕ'_B only consists of MMD

between empirical distributions, thus it can be implemented efficiently. From (2), we have that

$$\begin{aligned} & \left\| \mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}} - \left\| \mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}} \\ &= \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(\hat{x}_{0,i}, \hat{x}_{0,j}) \right. \\ & \quad \left. - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, \hat{x}_{0,j}) \right)^{1/2} \\ & \quad - \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(\hat{x}_{1,i}, \hat{x}_{1,j}) \right. \\ & \quad \left. - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, \hat{x}_{1,j}) \right)^{1/2}. \end{aligned} \quad (97)$$

This completes the proof. \square

APPENDIX G PROOF OF PROPOSITION 4

Proof. For any $P_0 \in \mathcal{P}_0$, we have that

$$\begin{aligned} & E_{P_0}[\phi_B(x^n)] \\ &= P_0 \left(\left\| \mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}} - \left\| \mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}} \geq 0 \right) \\ &= P_0 \left(\left\| \mu_{\hat{P}_n} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}}^2 - \left\| \mu_{\hat{P}_n} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}}^2 \geq 0 \right) \\ &= P_0 \left(E_{x \sim \hat{P}_n, x' \sim \hat{P}_n} [k(x, x')] + E_{y \sim \hat{Q}_m^0, y' \sim \hat{Q}_m^0} [k(y, y')] \right. \\ & \quad \left. - 2E_{x \sim \hat{P}_n, y \sim \hat{Q}_m^0} [k(x, y)] - E_{x \sim \hat{P}_n, x' \sim \hat{P}_n} [k(x, x')] \right. \\ & \quad \left. - E_{y \sim \hat{Q}_m^1, y' \sim \hat{Q}_m^1} [k(y, y')] + 2E_{x \sim \hat{P}_n, y \sim \hat{Q}_m^1} [k(x, y)] \geq 0 \right) \\ &= P_0 \left(-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m k(x_j, \hat{x}_{0,i}) + \frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m k(x_j, \hat{x}_{1,i}) \right. \\ & \quad \left. + E_{y \sim \hat{Q}_m^0, y' \sim \hat{Q}_m^0} [k(y, y')] - E_{y \sim \hat{Q}_m^1, y' \sim \hat{Q}_m^1} [k(y, y')] \geq 0 \right). \end{aligned} \quad (98)$$

We will then bound the error probability in (98) using the McDiarmid's inequality [66]. Define

$$\begin{aligned} & F(x_1, x_2, \dots, x_n) \\ &= -\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m k(x_j, \hat{x}_{0,i}) + \frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m k(x_j, \hat{x}_{1,i}). \end{aligned} \quad (99)$$

To apply the McDiarmid's inequality [66], we first need to bound

$$\begin{aligned} & \sup_{x_1, \dots, x_j, \dots, x_n, x'_j \in \mathcal{X}} \left(F(x_1, \dots, x_j, \dots, x_n) \right. \\ & \quad \left. - F(x_1, \dots, x'_j, \dots, x_n) \right), \forall 1 \leq j \leq n. \end{aligned} \quad (100)$$

It can be shown that

$$\begin{aligned} & \sup_{x_1, \dots, x_j, \dots, x_n, x'_j \in \mathcal{X}} \left(F(x_1, \dots, x_j, \dots, x_n) \right. \\ & \quad \left. - F(x_1, \dots, x'_j, \dots, x_n) \right) \\ &= \sup_{x_1, \dots, x_j, \dots, x_n, x'_j \in \mathcal{X}} \frac{2}{n} \left(\frac{1}{m} \sum_{i=1}^m (k(x'_j, \hat{x}_{0,i}) - k(x_j, \hat{x}_{0,i})) \right. \\ & \quad \left. + \frac{1}{m} \sum_{i=1}^m (k(x_j, \hat{x}_{1,i}) - k(x'_j, \hat{x}_{1,i})) \right) \\ &\leq \frac{4K}{n}, \end{aligned} \quad (101)$$

where the last inequality is due to the fact that the kernel $k(\cdot, \cdot)$ is bounded. We will then consider the expectation of $F(x_1, x_2, \dots, x_n)$. From similar steps in (98), it follows that

$$\begin{aligned} & E_{P_0} [F(x_1, x_2, \dots, x_n)] + E_{y \sim \hat{Q}_m^0, y' \sim \hat{Q}_m^0} [k(y, y')] \\ & \quad - E_{y \sim \hat{Q}_m^1, y' \sim \hat{Q}_m^1} [k(y, y')] \\ &= \left\| \mu_{P_0} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}}^2 - \left\| \mu_{P_0} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}}^2. \end{aligned} \quad (102)$$

Note that for any $P_0 \in \mathcal{P}_0$, we have that $\left\| \mu_{P_0} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}}^2 - \left\| \mu_{P_0} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}}^2 \leq 0$. It then follows that

$$\begin{aligned} & P_0 \left(-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m k(x_j, \hat{x}_{0,i}) + \frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m k(x_j, \hat{x}_{1,i}) \right. \\ & \quad \left. + E_{y \sim \hat{Q}_m^0, y' \sim \hat{Q}_m^0} [k(y, y')] - E_{y \sim \hat{Q}_m^1, y' \sim \hat{Q}_m^1} [k(y, y')] \geq 0 \right) \\ &= P_0 \left(-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m (k(x_j, \hat{x}_{0,i}) - k(x_j, \hat{x}_{1,i})) \right. \\ & \quad \left. - E_{P_0} \left[-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m (k(x_j, \hat{x}_{0,i}) - k(x_j, \hat{x}_{1,i})) \right] \right. \\ & \quad \left. \geq E_{y \sim \hat{Q}_m^1, y' \sim \hat{Q}_m^1} [k(y, y')] - E_{y \sim \hat{Q}_m^0, y' \sim \hat{Q}_m^0} [k(y, y')] \right. \\ & \quad \left. - E_{P_0} \left[-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m (k(x_j, \hat{x}_{0,i}) - k(x_j, \hat{x}_{1,i})) \right] \right) \\ &= P_0 \left(-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m (k(x_j, \hat{x}_{0,i}) - k(x_j, \hat{x}_{1,i})) \right. \\ & \quad \left. - E_{P_0} \left[-\frac{2}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m (k(x_j, \hat{x}_{0,i}) - k(x_j, \hat{x}_{1,i})) \right] \right. \\ & \quad \left. \geq \left\| \mu_{P_0} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}}^2 - \left\| \mu_{P_0} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}}^2 \right) \\ &\leq \exp \left(-\frac{n \left(\left\| \mu_{P_0} - \mu_{\hat{Q}_m^1} \right\|_{\mathcal{H}}^2 - \left\| \mu_{P_0} - \mu_{\hat{Q}_m^0} \right\|_{\mathcal{H}}^2 \right)^2}{8K^2} \right), \end{aligned} \quad (103)$$

where the last inequality is from the McDiarmid's inequality [66]. We then have that

$$\begin{aligned} & \sup_{P_0 \in \mathcal{P}_0} E_{P_0}[\phi_B(x^n)] \\ & \leq \sup_{P_0 \in \mathcal{P}_0} \exp \left(- \frac{n \left(\|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}}^2 - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 \right)^2}{8K^2} \right). \end{aligned} \quad (104)$$

Since $\|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}}^2 - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 \geq 0$ and the exponential function is monotonically increasing, the optimization problem on the right-hand side of (104) can be solved by solving $\inf_{P_0 \in \mathcal{P}_0} \|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}}^2 - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2$. We then have that

$$\begin{aligned} & \inf_{P_0 \in \mathcal{P}_0} \|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}}^2 - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 \\ & = \inf_{P_0 \in \mathcal{P}_0} \left(\|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \right) \\ & \quad \times \left(\|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \right). \end{aligned} \quad (105)$$

From the triangle inequality of MMD [26], we have that

$$\|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} + \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \geq \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}. \quad (106)$$

Moreover, since $\|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \leq \theta$, it can be shown that

$$\begin{aligned} & \|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}} - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \geq \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - 2\|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \\ & \geq \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - 2\theta. \end{aligned} \quad (107)$$

It then follows that

$$\begin{aligned} & \inf_{P_0 \in \mathcal{P}_0} \|\mu_{P_0} - \mu_{\hat{Q}_m^1}\|_{\mathcal{H}}^2 - \|\mu_{P_0} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 \\ & \geq \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \left(\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} - 2\theta \right). \end{aligned} \quad (108)$$

Therefore, we have that

$$\begin{aligned} & \sup_{P_0 \in \mathcal{P}_0} E_{P_0}[\phi_B(x^n)] \\ & \leq \exp \left(- \frac{n \left(\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}}^2 - 2\theta \|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\|_{\mathcal{H}} \right)^2}{8K^2} \right). \end{aligned} \quad (109)$$

Following the same idea as in the proof of (32), (33) can also be proved. This completes the proof. \square

APPENDIX H PROOF OF THEOREM 4

Proof. For any $P_0 \in \mathcal{P}_0$, we have that

$$P_0 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} > \gamma_n \right) \leq P_0 \left(\|\mu_{\hat{P}_n} - \mu_{P_0}\|_{\mathcal{H}} > \gamma_n \right).$$

Set $\gamma_n = \sqrt{2K/n}(1 + \sqrt{-\log \alpha})$. From Lemma 4 in Appendix A, we have that for any $P_0 \in \mathcal{P}_0$,

$$P_0 \left(\|\mu_{\hat{P}_n} - \mu_{P_0}\|_{\mathcal{H}} > \gamma_n \right) \leq \alpha. \quad (110)$$

We then have that

$$\sup_{P_0 \in \mathcal{P}_0} P_0 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} > \gamma_n \right) \leq \alpha. \quad (111)$$

Note that $\gamma_n \rightarrow 0$ as $n \rightarrow \infty$. For any $\gamma > 0$, there exists an integer n_0 such that $\gamma_n < \gamma$ for all $n > n_0$. We then have that for large n ,

$$\begin{aligned} & \left\{ P' \in \mathcal{P} : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma_n \right\} \\ & \subseteq \left\{ P' \in \mathcal{P} : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma \right\}. \end{aligned} \quad (112)$$

It then follows that

$$\begin{aligned} & \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_1 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \leq \gamma_n \right) \\ & \geq \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_1 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \leq \gamma \right) \\ & \geq \inf_{P_1 \in \mathcal{P}_1} \inf_{\{P' \in \mathcal{P} : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma\}} D(P' \| P_1), \end{aligned} \quad (113)$$

where the last inequality is from the Sanov's theorem [4] and the fact that $\{P' \in \mathcal{P} : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma\}$ is closed w.r.t. the weak topology. Let

$$\Gamma = \{P' \in \mathcal{P} : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma\}. \quad (114)$$

Since MMD metrizes the weak convergence on \mathcal{P} [45], [46], and KL divergence is lower semi-continuous with respect to the weak topology of \mathcal{P} (see Lemma 3 in Appendix A), we have that for any $\epsilon > 0$ and $P_1 \in \mathcal{P}_1$, there exists a neighborhood U of P_0 defined by MMD such that $D(P' \| P_1) \geq D(P_0 \| P_1) - \epsilon$ for any $P' \in U$.

Specifically, for any $P_0 \in \mathcal{P}_0$, define the neighborhood of P_0 with radius γ as $U(P_0, \gamma) = \{P \in \mathcal{P} : \|\mu_P - \mu_{P_0}\|_{\mathcal{H}} \leq \gamma\}$. From the lower semi-continuity of KL divergence, we have that for any $\epsilon > 0$ and $P_1 \in \mathcal{P}_1$, there exists $\gamma(\epsilon, P_0) > 0$ such that $D(P' \| P_1) \geq D(P_0 \| P_1) - \epsilon$ for any $P' \in U(P_0, \gamma(\epsilon, P_0))$. Therefore, for any $P' \in \bigcup_{P_0 \in \mathcal{P}_0} U(P_0, \gamma(\epsilon, P_0))$, there exists $P_0 \in \mathcal{P}_0$ such that $D(P' \| P_1) \geq D(P_0 \| P_1) - \epsilon$.

For a given $\epsilon > 0$ and $P_1 \in \mathcal{P}_1$, let $\gamma^* = \min_{P_0 \in \mathcal{P}_0} \gamma(\epsilon, P_0)$. Since $\gamma(\epsilon, P_0) > 0$ holds for any $P_0 \in \mathcal{P}_0$ and \mathcal{P}_0 is a closed set, we have that $\gamma^* > 0$. Let

$$\Gamma^* = \{P' \in \mathcal{P} : \inf_{P \in \mathcal{P}_0} \|\mu_{P'} - \mu_P\|_{\mathcal{H}} \leq \gamma^*\}. \quad (115)$$

We then have that

$$\Gamma^* = \bigcup_{P_0 \in \mathcal{P}_0} U(P_0, \gamma^*) \subseteq \bigcup_{P_0 \in \mathcal{P}_0} U(P_0, \gamma(\epsilon, P_0)). \quad (116)$$

We then have that for any $P' \in \Gamma^*$, there exists a $P_0 \in \mathcal{P}_0$ such that

$$D(P' \| P_1) \geq D(P_0 \| P_1) - \epsilon. \quad (117)$$

It then follows that there exists $P_0 \in \mathcal{P}_0$ such that

$$\inf_{P' \in \Gamma^*} D(P' \| P_1) \geq D(P_0 \| P_1) - \epsilon. \quad (118)$$

We then have that for any $\epsilon > 0$,

$$\inf_{P' \in \Gamma^*} D(P' \| P_1) \geq \inf_{P_0 \in \mathcal{P}_0} D(P_0 \| P_1) - \epsilon. \quad (119)$$

Since ϵ can be arbitrarily small, it then follows that

$$\begin{aligned} & \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_1 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \leq \gamma_n \right) \\ & \geq \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1). \end{aligned} \quad (120)$$

Combining (120) with Proposition 5, we have that

$$\begin{aligned} & \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1} [1 - \phi_N(x^n)] \\ & = \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log P_1 \left(\inf_{P \in \mathcal{P}_0} \|\mu_{\hat{P}_n} - \mu_P\|_{\mathcal{H}} \leq \gamma_n \right) \\ & = \sup_{\phi: P_F(\phi) \leq \alpha} \inf_{P_1 \in \mathcal{P}_1} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E_{P_1} [1 - \phi(x^n)] \\ & = \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0 \| P_1). \end{aligned} \quad (121)$$

This completes the proof. \square

ACKNOWLEDGMENT

The work of Z. Sun and S. Zou was supported in part by the National Science Foundation under Grants 1948165, 2106560 and 2112693.

REFERENCES

- [1] Z. Sun and S. Zou, "A data-driven approach to robust hypothesis testing using kernel MMD uncertainty sets," in *Proc. International Symposium on Information Theory (ISIT)*, pp. 3056–3061, 2021.
- [2] Z. Sun and S. Zou, "Robust hypothesis testing with kernel uncertainty sets," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 3309–3314, 2022.
- [3] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press, 2018.
- [4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: John Wiley & Sons, 2006.
- [5] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Prentice Hall PTR, 1993.
- [6] P. J. Huber, "A robust version of the probability ratio test," *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
- [7] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, 2009.
- [8] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, 2017.
- [9] R. Gao, L. Xie, Y. Xie, and H. Xu, "Robust hypothesis testing using Wasserstein uncertainty sets," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, pp. 7902–7912, 2018.
- [10] L. Xie, R. Gao, and Y. Xie, "Robust hypothesis testing with Wasserstein uncertainty sets," *arXiv preprint arXiv:2105.14348*, 2021.
- [11] J. Wang and Y. Xie, "A data-driven approach to robust hypothesis testing using Sinkhorn uncertainty sets," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 3315–3320, 2022.
- [12] M. Barni and B. Tondi, "The source identification game: An information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, 2013.
- [13] Y. Jin and L. Lai, "On the adversarial robustness of hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 69, pp. 515–530, 2021.
- [14] H. Rieder, "Least favorable pairs for special capacities," *The Annals of Statistics*, vol. 5, no. 5, pp. 909–921, 1977.
- [15] F. Österreicher, "On the construction of least favourable pairs of distributions," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 43, pp. 49–55, 1978.
- [16] T. Bednarski, "On solutions of minimax test problems for special capacities," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 58, pp. 397–405, 1981.
- [17] R. Hafner, "Simple construction of least favourable pairs of distributions and of robust tests for Prokhorov-neighbourhoods," *Series Statistics*, vol. 13, no. 1, pp. 33–46, 1982.
- [18] S. Kassam, "Robust hypothesis testing for bounded classes of probability densities (corresp.)," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 242–247, 1981.
- [19] R. Hafner, "Construction of minimax-tests for bounded families of probability-densities," *Metrika*, vol. 40, no. 1, pp. 1–23, 1993.
- [20] K. Vastola and H. Poor, "On the p-point uncertainty class (corresp.)," *IEEE Transactions on Information Theory*, vol. 30, no. 2, pp. 374–376, 1984.
- [21] M. Fauß, A. M. Zoubir, and H. V. Poor, "Minimax robust detection: Classic results and recent advances," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2252–2283, 2021.
- [22] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.
- [23] C. Pandit, S. Meyn, and V. Veeravalli, "Asymptotic robust Neyman-Pearson hypothesis testing based on moment classes," in *Proc. International Symposium on Information Theory (ISIT)*, p. 220, 2004.
- [24] A. Berline and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.
- [25] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Hilbert space embeddings and metrics on probability measures," *Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.
- [26] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.
- [27] Y. Altun and A. Smola, "Unifying divergence minimization and statistical inference via convex duality," in *Proc. Annual Conference on Learning Theory*, pp. 139–153, 2006.
- [28] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur, "Two-stage sampled Learning theory on distributions," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 38, pp. 948–957, 2015.
- [29] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton, "Fast two-sample testing with analytic representations of probability measures," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, p. 1981–1989, 2015.
- [30] K. Fukumizu, A. Gretton, G. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur, "Kernel choice and classifiability for RKHS embeddings of probability distributions," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 22, pp. 1750–1758, 2009.
- [31] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, "A fast, consistent kernel two-sample test," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 22, pp. 673–681, 2009.
- [32] 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 *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1205–1213, 2012.
- [33] D. J. Sutherland, H. Tung, H. Strathmann, S. De, A. Ramdas, A. J. Smola, and A. Gretton, "Generative models and model criticism via optimized maximum mean discrepancy," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
- [34] W. Zaremba, A. Gretton, and M. Blaschko, "B-test: A non-parametric, low variance kernel two-sample test," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 26, pp. 755–763, 2013.
- [35] S. Zhu, B. Chen, Z. Chen, and P. Yang, "Asymptotically optimal one- and two-sample testing with kernels," *IEEE Transactions on Information Theory*, vol. 67, no. 4, pp. 2074–2092, 2021.
- [36] J. R. Lloyd and Z. Ghahramani, "Statistical model criticism using kernel two sample tests," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 829–837, 2015.
- [37] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for Interpretability," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 29, pp. 2288–2296, 2016.
- [38] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data streams," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5785–5797, 2017.
- [39] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of geometric structures over networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 19, pp. 5034–5046, 2017.
- [40] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 3366–3374, 2015.
- [41] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf, "Kernel distributionally robust optimization: Generalized duality theorem and stochastic

- approximation,” in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 130, pp. 280–288, 2021.
- [42] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg, 2009.
- [43] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, “Kernel mean embedding of distributions: A review and beyond,” *Foundations and Trends in Machine Learning*, vol. 10, no. 1–2, pp. 1–144, 2017.
- [44] 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*, vol. 11, no. 50, pp. 1517–1561, 2010.
- [45] C.-J. Simon-Gabriel and B. Schölkopf, “Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions,” *Journal of Machine Learning Research*, vol. 19, no. 44, pp. 1–29, 2018.
- [46] B. Sriperumbudur, “On the optimal estimation of probability measures in weak and strong topologies,” *Bernoulli*, vol. 22, no. 3, pp. 1839–1893, 2016.
- [47] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.2.” <http://cvxr.com/cvx>, Jan. 2020.
- [48] A. Tychonoff, “Über die topologische erweiterung von räumen,” *Mathematische Annalen*, vol. 102, pp. 544–561, 1930.
- [49] P. Johnstone, “Tychonoff’s theorem without the axiom of choice,” *Fundamenta Mathematicae*, vol. 113, no. 1, pp. 21–35, 1981.
- [50] M. Sion, “On general minimax theorems,” *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171 – 176, 1958.
- [51] J. Vert, K. Tsuda, and B. Schölkopf, *A Primer on Kernel Methods*, pp. 35–70. Cambridge, MA, USA: MIT Press, 2004.
- [52] J. L. W. V. Jensen, “Sur les fonctions convexes et les inégalités entre les valeurs moyennes,” *Acta Mathematica*, vol. 30, no. none, pp. 175 – 193, 1906.
- [53] A. W. Roberts and D. E. Varberg, “Another proof that convex functions are locally lipschitz,” *The American Mathematical Monthly*, vol. 81, no. 9, pp. 1014–1016, 1974.
- [54] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.
- [55] J. W. Lockhart, G. M. Weiss, J. C. Xue, S. T. Gallagher, A. B. Grosner, and T. T. Pulickal, “Design considerations for the WISDM smart phone-based sensor mining architecture,” in *Proc. International Workshop on Knowledge Discovery from Sensor Data*, pp. 25–33, 2011.
- [56] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers,” *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, p. 74–82, 2011.
- [57] G. M. Weiss and J. W. Lockhart, “The impact of personalization on smartphone-based activity recognition,” in *Proc. AAAI Workshop on Activity Context Representation: Techniques and Languages*, pp. 98–104, 2012.
- [58] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [59] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman, “On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions,” in *Proc. AAAI Conference on Artificial Intelligence*, p. 3571–3577, 2015.
- [60] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, 1965.
- [61] T. V. Erven and P. Harremoës, “Rényi divergence and Kullback-Leibler divergence,” *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.
- [62] E. Posner, “Random coding strategies for minimum entropy,” *IEEE Transactions on Information Theory*, vol. 21, no. 4, pp. 388–391, 1975.
- [63] M. S. Halmos, *Measure Theory*. New York, NY, USA: Springer-Verlag, 1950.
- [64] A. Klenke, *Probability Theory*. Springer-Verlag London, 2014.
- [65] Y. V. Prokhorov, “Convergence of random processes and limit theorems in probability theory,” *Theory of Probability & Its Applications*, vol. 1, no. 2, pp. 157–214, 1956.
- [66] C. McDiarmid, “On the method of bounded differences,” in *Surveys in combinatorics*, pp. 148–188, Cambridge University Press, 1989.

received the B.S. degree from Beijing Institute of Technology, Beijing, China in 2019. His research interests are on hypothesis testing, quickest change detection, distributionally robust optimization and reinforcement learning.

Shaofeng Zou (S’14-M’16) is an Assistant Professor, at the Department of Electrical Engineering, University at Buffalo, the State University of New York. He received the Ph.D. degree in Electrical and Computer Engineering from Syracuse University in 2016. He received the B.E. degree (with honors) from Shanghai Jiao Tong University, Shanghai, China, in 2011. He was a postdoctoral research associate at the Coordinated Science Lab, University of Illinois at Urbana-Champaign during 2016–2018. Dr. Zou’s research interests include reinforcement learning, machine learning, statistical signal processing and information theory. He received the National Science Foundation CRII award in 2019 and the 2023 AAAI Distinguished Paper Award.

Zhongchang Sun (S’20) is a PhD student at the Department of Electrical Engineering, University at Buffalo, the State University of New York. He