Robust Hypothesis Testing with Kernel Uncertainty Sets

Zhongchang Sun and Shaofeng Zou
Electrical Engineering
University at Buffalo, the State University of New York
Buffalo, NY, USA
Email: zhongcha@buffalo.edu, szou3@buffalo.edu

***Abstract*—In this paper, the robust hypothesis testing problem is investigated, where under the null and the alternative hypotheses, the distributions are assumed to be in some uncertainty sets. The uncertainty sets are constructed in a data-driven manner, i.e., they are centered around empirical distributions. The distance between kernel mean embeddings of distributions in the reproducing kernel Hilbert space is used as the distance metric of uncertainty sets. The Bayesian setting is studied, where the goal is to minimize the worst-case error probability. An optimal test is firstly obtained for the case with a finite alphabet. For the case with an infinite alphabet, a tractable approximation is proposed to quantify the worst-case error probability, and a kernel smoothing method is further applied to design test that generalizes to unseen samples. A heuristic robust kernel test is also proposed and proved to be exponentially consistent. Numerical results are provided to demonstrate the performance of the proposed tests.**

***Index Terms*—Bayesian setting, kernel smoothing, worst-case error quantification, kernel robust test**

I. Introduction

Hypothesis testing problem has been widely studied where the goal is to distinguish among different hypotheses with a small probability of error [1]–[3]. For simple hypothesis testing problems where samples under each hypothesis follow a fixed and known distribution, the likelihood ratio test is optimal under various settings, e.g., the Neyman-Pearson setting and the Bayesian setting. However, the likelihood ratio test requires exact knowledge of the data-generating distributions. When the distributions in the likelihood ratio test deviate from the true data-generating distributions, the performance may degrade significantly. To address this problem, robust hypothesis testing is studied, e.g., [4]–[18], where the true distributions belong to some uncertainty sets of distributions, which are centered around 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.

One of the earliest work on robust hypothesis testing dates back to Huber [4], where a censored version of the nominal likelihood ratio test was proposed for ϵ-contamination uncertainty sets and the total-variation uncertainty sets. A work by Levy [5] considered the uncertainty sets defined via Kullback-Leibler (KL) divergence, and proposed the likelihood ratio test based on the least-favorable distributions (LFDs). In [7], the robust hypothesis testing problem under the Bernoulli distribution was investigated. In [8], the uncertainty sets were constructed via distortion constraints. These works mainly focus on the 1-dimensional setting, where nominal distributions can be estimated from samples. However, when it comes to the high-dimensional setting, estimating distributions accurately from historical data is difficult, and hence the existing approaches may not be applicable anymore. In this paper, we propose a data-driven approach to construct the nominal distributions. We use the empirical distributions of samples from the null and alternative hypotheses as the nominal distributions directly. We note that in this case, the uncertainty sets defined via KL divergence is not applicable since such uncertainty sets only contain distributions supported on the training samples, which may be problematic if the alphabet is actually infinite.

The data-driven approach has been studied in [17], [18], where uncertainty sets are centered around empirical distributions via the Wasserstein distance. A nearly-optimal detector and an optimal test were derived in [17] and [18], respectively. However, Wasserstein distance based approach has certain drawbacks. Firstly, the Wasserstein distance between the empirical distribution with m samples and its data-generating distribution is bounded by $\mathcal{O}(m^{-1/d})$ [19], which depends on the dimension d of the data. Therefore, when used to choose radii of uncertainty sets to guarantee that the data-generating distributions lie in the uncertainty sets with high probability, it might be too pessimistic when d is large. Moreover, coefficients in such a concentration bound depend on the true distribution [18] which is unknown, thus this makes it difficult to use in practice. Secondly, 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 [20], the uncertainty sets are constructed using moment classes, where a finite alphabet was considered and an asymptotically optimal test was designed. The moment uncertainty sets in [20] 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) [21]–[23] 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, and we further take the supremum of g over an RKHS to account for the worst case. Then this leads to uncertainty sets centered at $\hat{P}$ and defined by MMD. Compared with the Wasserstein distance, the kernel MMD between the empirical distribution with m samples and the population distribution can be bounded by $\mathcal{O}(1/\sqrt{m})$, which is dimension-free and also does not depend on the data-generating distribution. Moreover, the kernel MMD is computationally efficient to evaluate.

In this paper, we focus on the Bayesian setting where the goal is to minimize the worst-case error probability, which is different from the Neyman-Pearson setting considered in [24]. Moreover, this paper focuses on the probability of error using approaches based on LFDs, which is different from the asymptotic approach in [24] that analyzes the error exponent. We first study the case with a finite alphabet and obtain the optimal test via the strong duality of kernel robust optimization. For the case with an infinite alphabet, we propose a tractable approximation to quantify the worst-case error probability, and then apply the kernel smoothing method to design a robust test that generalizes to unseen data. We also propose a heuristic robust kernel test and show that it is exponentially consistent.

II. Preliminaries: Maximum Mean Discrepancy

We first give a brief introduction to idea of kernel mean embedding and the MMD [21], [22]. Let $\mathcal{H}$ denote the RKHS associated with a kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. There exists a feature map $k(x,\cdot) : \mathcal{X} \to \mathcal{H}$ such that $k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}}$ defines an inner product on $\mathcal{H}$. The RKHS $\mathcal{H}$ is equipped with a reproducing property such that $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}, x \in \mathcal{X}$. The MMD between two distributions P_0 and P_1 is defined as

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

The kernel mean embedding of a distribution P is defined as $\mu_P = \int k(x,\cdot)dP$. With the reproducing property of $\mathcal{H}$, 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 [23]:

$$d_{\text{MMD}}(P_0,P_1) = \left\|\mu_{P_0} - \mu_{P_1}\right\|_{\mathcal{H}}, \quad (2)$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm on $\mathcal{H}$. If a kernel k is characteristic [25], $d_{\text{MMD}}(\cdot,\cdot)$ is a metric on $\mathcal{P}$ [23], [26]. In this paper, we consider kernels such that the weak convergence on $\mathcal{P}$ can be metrized by MMD [27], [28], e.g., Gaussian kernels and Laplacian kernels.

III. 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 supported on $\mathcal{X}$. For a simple hypothesis testing problem, the goal is to distinguish between the null hypothesis that the sample x is generated from $P_0 \in \mathcal{P}$ and the alternative hypothesis that the sample x is generated from $P_1 \in \mathcal{P}$. For a randomized test $\phi : \mathcal{X} \to [0,1]$, it accepts the null hypothesis H_0 with probability $1-\phi(x)$ and accepts the alternative hypothesis H_1 with probability $\phi(x)$. The error probability in the Bayesian setting with equal prior is given by

$$P_E(\phi) \triangleq \frac{1}{2}E_{P_0}\big[\phi(x)\big] + \frac{1}{2}E_{P_1}\big[1-\phi(x)\big]. \quad (3)$$

In this paper, we consider a data-driven setting where P_0, P_1 are unknown, and only training samples from them are available. Suppose we have two sequences of training samples: $\hat{x}_0^m = (\hat{x}_{0,1}, \hat{x}_{0,2}, \cdots, \hat{x}_{0,m})$ and $\hat{x}_1^m = (\hat{x}_{1,1}, \hat{x}_{1,2}, \cdots, \hat{x}_{1,m})$ generated from P_0, P_1, respectively, where m is the training sample size. Let $\hat{Q}_m^l = \frac{1}{m}\sum_{i=1}^m \delta_{\hat{x}_{l,i}}$ be the empirical distributions of $\hat{x}_l^m$ for $l = 0, 1$, where $\delta_{\hat{x}_{l,i}}$ denotes the Dirac measure on $\hat{x}_{l,i}$. To model the uncertainty of P_0, P_1, we define two uncertainty sets centered around the empirical distributions $\hat{Q}_m^l, l = 0, 1$, respectively. Specifically, we define the uncertainty sets via the MMD:

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

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

In [20], the moment class is defined as $\{P : E_P[f] \le \theta\}$, where f is a real-valued 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 of g such that $\|g\|_{\mathcal{H}} \le 1$ over 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.

For the Bayesian robust hypothesis testing, the goal is to solve the following problem:

$$\inf_{\phi} \sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1} P_E(\phi). \quad (5)$$

IV. Finite-Alphabet Robust Hypothesis Testing

In this section, we focus on the case with a finite alphabet, i.e., $|\mathcal{X}| < \infty$. A finite-dimensional version of (5) is considered, and a minimax optimal test is derived. The results in this finite alphabet setting will be useful for the infinite alphabet setting in the next section.

Let $\mathcal{X} = \{z_1, z_2, \cdots, z_N\}$, $N = |\mathcal{X}|$ and $\hat{x}_{l,j}$ $(l = 0, 1) \in \{z_i\}_{i=1}^N$ for $j = 1, \cdots, m$. In this case, $P_E(\phi) = \frac{1}{2}\sum_{i=1}^N (1 - \phi_N(z_i))P_1^N(z_i) + \phi_N(z_i)P_0^N(z_i)$, where we introduce the superscript N on P_0 and P_1 to emphasize its dependence on N, and therefore, (5) can be written as

$$\frac{1}{2}\min_{\phi_N} \sup_{P_0^N\in\mathcal{P}_0, P_1^N\in\mathcal{P}_1} \sum_{i=1}^N (1-\phi_N(z_i))P_1^N(z_i) + \phi_N(z_i)P_0^N(z_i). \quad (6)$$

Since (6) is a minimax problem and thus cannot be solved directly. We reformulate it equivalently as a finite-dimensional convex optimization problem in the following Theorem.

Theorem 1. *The minimax problem in* (6) *is equivalent to*

$$\begin{aligned}\frac{1}{2}\min_{\substack{\phi_N, f_0, g_0 \in \mathbb{R} \\ f_1, g_1 \in \mathcal{H}}} & f_0 + g_0 + \frac{1}{m}\sum_{i=1}^{m} f_1(\hat{x}_{1,i}) + \frac{1}{m}\sum_{i=1}^{m} g_1(\hat{x}_{0,i}) \\ & + \theta\|f_1\|_{\mathcal{H}} + \theta\|g_1\|_{\mathcal{H}} \\ \textit{subject to } & 1 - \phi_N(z_i) \le f_0 + f_1(z_i) \textit{ for } i = 1, \cdots, N \\ & \phi_N(z_i) \le g_0 + g_1(z_i) \textit{ for } i = 1, \cdots, N \\ & 0 \le \phi_N(z_i) \le 1 \textit{ for } i = 1, \cdots, N, \end{aligned} \tag{7}$$

which is a finite-dimensional convex optimization problem.

Proof. From the strong duality of kernel robust optimization [29], we have that (7) is equivalent to (6). From the robust representer theorem [29], the functions f_1, g_1 admit the finite expansions $f_1(\cdot) = \sum_{i=1}^{N} \alpha_i k(z_i, \cdot)$ and $g_1(\cdot) = \sum_{i=1}^{N} \beta_i k(z_i, \cdot)$. Therefore, the optimization problem in (6) can be reformulated as a finite-dimensional convex optimization problem thus is tractable in practice. □

By solving (7), 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 (6) by plugging ϕ_N^* back to (6).

V. Infinite-Alphabet Robust Hypothesis Testing

In this section, we consider the case where $\mathcal{X}$ is infinite. We first propose a tractable approximation to quantify the worst-case error probability of (5). This tractable approximation builds a connection between the finite-alphabet case and the infinite-alphabet case. We then design a robust test for (5) by extending the optimal test of the case with a finite alphabet to the case with an infinite alphabet via the kernel smoothing method. Finally, we propose another heuristic robust kernel test which is further shown to be exponentially consistent under the Bayesian setting.

A. Worst-Case Error Probability Quantification

For the infinite-alphabet case, (7) is infinite-dimensional, thus is intractable. To simplify the analysis of (5), we first interchange the sup and inf operators in (5) 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 solving the maximization problem. In the following, we use capital letter P to denote a distribution and lower case letter p to denote its probability density (mass) function.

Proposition 1. *The minimax problem in* (5) *has the following reformulation:*

$$\begin{aligned}&\inf_{\phi}\sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1} P_E(\phi) = \sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1}\inf_{\phi} P_E(\phi) \\ =&\frac{1}{2}\sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1}\int \min\big\{p_0(x), p_1(x)\big\}dx. \end{aligned} \tag{8}$$

Proof. The error probability $P_E(\phi)$ is continuous, real-valued and linear in ϕ, P_0 and P_1. For any distributions $Q_1, Q_2 \in \mathcal{P}_l, l \in \{0,1\}$, from the triangle inequality of MMD [23], the convex combination $\lambda Q_1 + (1-\lambda)Q_2, 0 < \lambda < 1$, lies in $\mathcal{P}_l$. Therefore, the uncertainty sets $\mathcal{P}_0$ 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 set of $[0,1]$. Since $\mathcal{X}$ is compact, from the Tychonoff's theorem [30], [31], Φ is compact with respect to the product topology. Moreover, for any $\phi_1, \phi_2 \in \Phi$, the convex combination $\lambda\phi_1 + (1-\lambda)\phi_2, 0 < \lambda < 1$ lies in Φ. Therefore, Φ is convex. From the Sion's minimax theorem [32], we have that

$$\begin{aligned}&\inf_{\phi}\sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1} P_E(\phi) = \sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1}\inf_{\phi} P_E(\phi) \\ =& \sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1}\frac{1}{2}\int \mathbb{I}_{\left\{\frac{p_1(x)}{p_0(x)}\ge 1\right\}} p_0(x)dx \\ &+ \frac{1}{2}\int \mathbb{I}_{\left\{\frac{p_1(x)}{p_0(x)}< 1\right\}} p_1(x)dx \\ =&\frac{1}{2}\sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1}\int \min\big\{p_0(x), p_1(x)\big\}dx, \end{aligned} \tag{9}$$

where $\mathbb{I}$ denote 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. □

Observe that the problem in (8) 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 (8). With this tractable approximation, the worst-case error probability in (8) 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 a distribution supported on the whole space $\mathcal{X}$ and $\{z_i\}_{i=1}^N$ be N samples generated from P. We propose the following approximation of (8)

$$\frac{1}{2}\sup_{P_0^N\in\mathcal{P}_0^N, P_1^N\in\mathcal{P}_1^N}\ \sum_{i=1}^{N}\min\big\{p_0^N(z_i), p_1^N(z_i)\big\}, \tag{10}$$

where $\mathcal{P}_l^N (l = 0, 1)$ denotes the collection of distributions that are supported on $\{z_i\}_{i=1}^N$ and satisfy $\left\|\mu_{P_l^N} - \mu_{\hat{Q}_m^l}\right\|_{\mathcal{H}} \le \theta$. We note that (10) is a finite-dimensional convex optimization problem which can be solved by standard optimization tools. Let $f(\mathcal{P}_0, \mathcal{P}_1) = \frac{1}{2}\sup_{P_0\in\mathcal{P}_0, P_1\in\mathcal{P}_1}\int \min\big\{p_0(x), p_1(x)\big\}dx$ and $f(\mathcal{P}_0^N, \mathcal{P}_1^N) = \frac{1}{2}\sup_{P_0^N\in\mathcal{P}_0^N, P_1^N\in\mathcal{P}_1^N}\sum_{i=1}^{N}\min\big\{p_0^N(z_i), p_1^N(z_i)\big\}$. The following theorem demonstrates the relationship between the problem (8) and its tractable approximation (10).

Theorem 2. *As $N \to \infty$, $f(\mathcal{P}_0^N, \mathcal{P}_1^N)$ converges to $f(\mathcal{P}_0, \mathcal{P}_1)$ almost surely.*

To prove Theorem 2, we show that $\int \min\big\{p_0(x), p_1(x)\big\}dx$ is upper semi-continuous in P_0, P_1 with respect to the weak convergence in the following lemma.

Lemma 1. *$\int \min\big\{p_0(x), p_1(x)\big\}dx$ is upper semi-continuous in P_0, P_1 with respect to the weak convergence.*

Proof sketch. To prove Lemma 1, we first show that $\int \min\big\{p_0(x), p_1(x)\big\}dx$ is concave with respect to P_0, P_1.

Let B be the σ-field on $\mathcal{X}$. Let $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_{|\mathcal{A}|}\}$ be the 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 $i = 1, 2, \cdots, |\mathcal{A}|$. We can 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 then the upper semi-continuity follows. □

Though the optimal solutions P_0^*, P_1^* cannot be derived from (8), (10) provides a lower bound on the worst-case error probability, and is asymptotically accurate as $N \to \infty$.

B. Robust Test via Kernel Smoothing

The optimal solution P_0^*, P_1^* of (8) are difficult to derive thus the likelihood ratio test between P_0^* and P_1^* is not applicable for our problem. Since $(P_0^{*,N}, P_1^{*,N}, \phi_N^*)$ is an optimal solution to (6), $P_0^{*,N}, P_1^{*,N}$ are optimal solutions to (10). In this section, we propose a kernel smoothing method to design a robust test that generalizes to the entire alphabet based on the following proposition.

Proposition 2. *There exists a subsequence of $\{P_0^{*,N}, P_1^{*,N}\}_{N=1}^{\infty}$ that converges weakly to an optimal solution of* (8)*.*

Proof. Observe that for any N, $\{P_0^{*,N}, P_1^{*,N}\}$ lies in the compact set $\mathcal{P}_0 \times \mathcal{P}_1$. Therefore, there exists a subsequence of $\{P_0^{*,N}, P_1^{*,N}\}_{N=1}^{\infty}$ that converges and the limit lies in $\mathcal{P}_0 \times \mathcal{P}_1$. Denote the sequence by $\{P_0^{*,N_t}, P_1^{*,N_t}\}_{t=1}^{\infty}$. Suppose $\{P_0^{*,N_t}, P_1^{*,N_t}\}_{t=1}^{\infty}$ converges weakly to $\{P_0', P_1'\}$. Since $P_0^{*,N}, P_1^{*,N}$ are optimal solutions to (10), we have that

$$\int \min\{p_0^*(x), p_1^*(x)\}dx = \lim_{t\to\infty} \int \min\{p_0^{*,N_t}, p_1^{*,N_t}\}dx \leq \int \min\{p_0'(x), p_1'(x)\}dx, \tag{11}$$

where the inequality is due to the upper semi-continuity of $\int \min\{p_0(x), p_1(x)\}dx$. Since P_0^*, P_1^* are optimal solutions of (8) and $P_0' \in \mathcal{P}_0, P_1' \in \mathcal{P}_1$, from (11), P_0', P_1' are optimal solutions of (8). This completes the proof. □

Note that $P_0^{*,N}, P_1^{*,N}$ are convex combinations of Dirac measures, from Proposition 2, we can extend them to the whole space via kernel smoothing to approximate P_0^*, P_1^*, i.e.,

$$\widetilde{P}_0^*(x) = \sum_{i=1}^{N} P_0^{*,N}(z_i)k(x, z_i),$$
$$\widetilde{P}_1^*(x) = \sum_{i=1}^{N} P_1^{*,N}(z_i)k(x, z_i). \tag{12}$$

The kernel functions have various choices. For example, the Gaussian kernel with bandwidth parameter σ: $k(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{\|x-y\|^2}{2\sigma^2}\big)$. After kernel smoothing, we define the likelihood ratio test $\tilde{\phi}$ between $\widetilde{P}_1^*(x)$ and $\widetilde{P}_0^*(x)$ over the whole space $\mathcal{X}$ to approximate the optimal test. The numerical results in Section VI show that $\tilde{\phi}$ performs well in practice.

C. A Heuristic Robust Kernel Test

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

Motivated by the fact that MMD can be used to measure the distance between distributions when the kernel k is characteristic, we propose the following test:

$$\phi_B(x^n) = \begin{cases} 1, & \text{if } S(x^n) \geq \gamma \\ 0, & \text{if } S(x^n) < \gamma, \end{cases} \tag{13}$$

where $S(x^n) = \inf_{P\in\mathcal{P}_0} \big\|\mu_{\hat{P}_n} - \mu_P\big\|_{\mathcal{H}} - \inf_{P\in\mathcal{P}_1} \big\|\mu_{\hat{P}_n} - \mu_P\big\|_{\mathcal{H}}$ and γ is a pre-specified threshold. The motivation for this test is as follows. We use "inf" to tackle the uncertainty of distributions and compare the closest distance between the empirical distribution of samples and two uncertainty sets. In the following theorem, we show that with the proper choice of γ, ϕ_B is exponentially consistent and can be implemented efficiently with a computational complexity of $\mathcal{O}\big(m^2 + n^2\big)$.

Theorem 3. *If $\gamma \in \Big(-\big\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\big\|_{\mathcal{H}} + 2\theta, \big\|\mu_{\hat{Q}_m^0} - \mu_{\hat{Q}_m^1}\big\|_{\mathcal{H}} - 2\theta\Big)$, ϕ_B is exponentially consistent, i.e.,*

$$\lim_{n\to\infty} -\frac{1}{n}\log P_0\big\{x^n : \phi_B(x^n) = 1\big\} \geq \inf_{P'\in\Gamma_0} D(P'\|P_0) > 0,$$
$$\lim_{n\to\infty} -\frac{1}{n}\log P_1\big\{x^n : \phi_B(x^n) = 0\big\} \geq \inf_{P'\in\Gamma_1} D(P'\|P_1) > 0.$$

where $\Gamma_0 = \Big\{P' : \inf_{P\in\mathcal{P}_0} \big\|\mu_{P'} - \mu_P\big\|_{\mathcal{H}} - \inf_{P\in\mathcal{P}_1} \big\|\mu_{P'} - \mu_P\big\|_{\mathcal{H}} \geq \gamma\Big\}$ and $\Gamma_1 = \Gamma_0^c$.

Moreover, ϕ_B can be equivalently written as

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

and can be implemented with a complexity of $\mathcal{O}\big(m^2 + n^2\big)$.

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 the form of an optimization problem and do not have 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 3. *For the heuristic robust kernel test in* (13)*, when $\gamma = 0$, we have that for $l = 0, 1$,*

$$\sup_{P_l\in\mathcal{P}_l} P_l\big\{x^n : \phi_B(x^n) = 1 - l\big\} \leq \exp\left(-\frac{n\Big(\big\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\big\|_{\mathcal{H}}^2 - 2\theta\big\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\big\|_{\mathcal{H}}\Big)^2}{8K^2}\right).$$

In Proposition 3, 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 rate

$\left(\left\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\right\|_{\mathcal{H}}^2 - 2\theta\left\|\mu_{\hat{Q}_m^1} - \mu_{\hat{Q}_m^0}\right\|_{\mathcal{H}}\right)^2/8K^2$, 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 of error probabilities will increase with radius θ. Proposition 3 provides a closed-form upper bound of the worst-case error probability in the non-asymptotic regime. In practice, this upper bound can be used to evaluate the worst-case risk of implementing ϕ_B. Moreover, combining Theorem 2 and Theorem 3, the performance gap between ϕ_B and the optimal test can be approximated.

VI. Simulation Results

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

We compare the performance of our kernel smoothing robust test $\tilde{\phi}$ and the heuristic robust kernel test ϕ_B. We first compare their performance under the multi-variate Gaussian distributions. We use 50 samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and 50 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. 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 (6). We then 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 our kernel smoothing robust test has a better performance than the heuristic robust kernel test. Moreover, with the increasing of sample size n, the error probabilities of the heuristic robust kernel test and the kernel smoothing robust test decay exponentially fast, which validates the theoretical result that the heuristic robust kernel test is exponentially consistent.

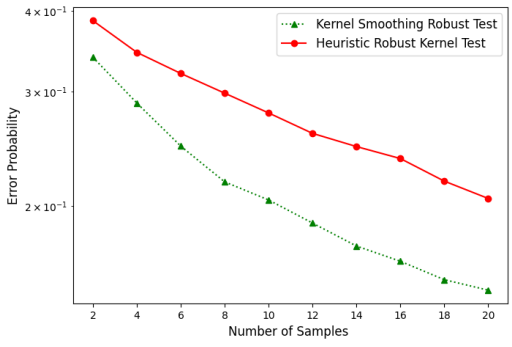

Fig. 1. Comparison of Two Tests

We then compare the performance of the two tests on a real data set. The dataset was released by the Wireless Sensor Data Mining (WISDM) Lab in October 2013, which was collected with the Actitracker system [33]–[35]. 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 jogging data from the person indexed by 685 and the walking data from the person indexed by 669 to form H_0 and H_1 respectively. A small portion of the data is used to construct the uncertainty sets. The radius θ of the uncertainty sets is chosen to be 0.03. We plot the log scale error probability as a function of testing sample size n. In Fig. 2, we have that the performance of the kernel smoothing robust test is better than the heuristic robust kernel test. Moreover, from Fig. 2, it can be seen that the error probabilities of the kernel smoothing robust test and the heuristic robust kernel test decay exponentially fast with sample size n, which validates our theoretical results.

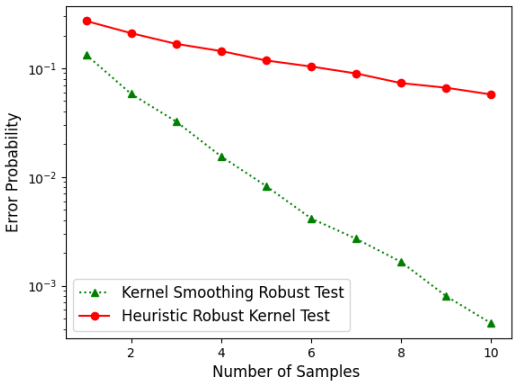

Fig. 2. Comparison of Two Tests in Real Data

VII. 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. We investigated the Bayesian setting where the goal is to minimize the worst-case error probability. We developed an approach to find 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, and we developed a kernel smoothing method to generalize to unseen data in the alphabet. We also developed a heuristic robust kernel test which was further shown to be exponentially consistent. The exact optimal solution for the infinite-alphabet case is challenging, and is of future interest.

Acknowledgment

The work of Z. Sun and S. Zou is partially supported by National Science Foundation (NSF) under grants CCF-1948165, CCF-2106560, and ECCS-2112693.

REFERENCES

[1] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*, Cambridge University Press, 2018.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 2006.

[3] S. M. Kay, *Fundamentals of Statistical Signal Processing*, Prentice Hall PTR, 1993.

[4] P. J. Huber, "A robust version of the probability ratio test," *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[5] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, 2009.

[6] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, 2017.

[7] 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.

[8] Y. Jin and L. Lai, "On the adversarial robustness of hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 69, pp. 515–530, 2021.

[9] H. Rieder, "Least favorable pairs for special capacities," *The Annals of Statistics*, vol. 5, no. 5, pp. 909–921, 1977.

[10] F. Österreicher, "On the construction of least favourable pairs of distributions," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 43, pp. 49–55, 1978.

[11] T Bednarski, "On solutions of minimax test problems for special capacities," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 58, pp. 397–405, 1981.

[12] 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.

[13] V.V. Veeravalli, T. Basar, and H.V. Poor, "Minimax robust decentralized detection," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 35–40, 1994.

[14] 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.

[15] R. Hafner, "Construction of minimax-tests for bounded families of probability-densities," *Metrika*, vol. 40, no. 1, pp. 1–23, 1993.

[16] 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.

[17] R. Gao, L. Xie, Y. Xie, and H. Xu, "Robust hypothesis testing using Wasserstein uncertainty sets," in *Proc. Advances Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7902–7912.

[18] L. Xie, R. Gao, and Y. Xie, "Robust hypothesis testing with Wasserstein uncertainty sets," *arXiv preprint arXiv:2105.14348*, 2021.

[19] 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.

[20] 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)*, 2004, p. 220.

[21] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, 2004.

[22] 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.

[23] 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.

[24] 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)*, 2021, pp. 3056–3061.

[25] 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.

[26] 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.

[27] 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.

[28] B. Sriperumbudur, "On the optimal estimation of probability measures in weak and strong topologies," *Bernoulli*, vol. 22, no. 3, pp. 1839–1893, 2016.

[29] 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 Artifical Intelligence and Statistics (AISTATS)*, 2021, vol. 130, pp. 280–288.

[30] A. Tychonoff, "Über die topologische erweiterung von räumen," *Mathematische Annalen*, vol. 102, pp. 544–561, 1930.

[31] P. Johnstone, "Tychonoff's theorem without the axiom of choice," *Fundamenta Mathematicae*, vol. 113, no. 1, pp. 21–35, 1981.

[32] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171 – 176, 1958.

[33] 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*, 2011, pp. 25–33.

[34] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[35] 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*, 2012, pp. 98–104.