Nonparametric Composite Hypothesis Testing in an Asymptotic Regime

Qunwei Li, *Student Member, IEEE*, Tiexing Wang, Donald J. Bucci, Yingbin Liang, *Senior Member, IEEE*, Biao Chen, *Fellow, IEEE*, and Pramod K. Varshney, *Life Fellow, IEEE*

***Abstract*—We investigate the nonparametric, composite hypothesis testing problem for arbitrary unknown distributions in the asymptotic regime where both the sample size and the number of hypothesis grow exponentially large. Such asymptotic analysis is important in many practical problems, where the number of variations that can exist within a family of distributions can be countably infinite. We introduce the notion of *discrimination capacity*, which captures the largest exponential growth rate of the number of hypothesis relative to the sample size so that there exists a test with asymptotically vanishing probability of error. Our approach is based on various distributional distance metrics in order to incorporate the generative model of the data. We provide analyses of the error exponent using the maximum mean discrepancy and Kolmogorov–Smirnov distance and characterize the corresponding discrimination rates, i.e., lower bounds on the discrimination capacity, for these tests. Finally, an upper bound on the discrimination capacity based on Fano's inequality is developed. Numerical results are presented to validate the theoretical results.**

***Index Terms*—Channel coding, maximum mean discrepancy, Kolmogorov–Smirnov distance, error exponent, discrimination rate.**

I. Introduction

INFORMATION theory was largely developed in the context of communication systems, where it plays an important role in characterizing performance limits. However, another important area where information theory has proved useful is in statistical inference, e.g., hypothesis testing. For *parametric* hypothesis testing problems, information theoretic tools such as joint typicality, the equipartition property, and Sanov's theorem have been developed to characterize the error exponent [1]–[3]. Information theory has also been applied to investigate a class of *parametric* hypothesis testing problems [4], [5], where correlated data samples are observed over multiple terminals and data compression needs to be carried out in a decentralized manner. Additionally, information theory has also been applied to solve *nonparametric* hypothesis testing problems under the Neyman-Pearson framework [6], [7].

In this paper, we apply information theoretic tools to study the nonparametric hypothesis testing problem, but with a focus on the average error probability instead of the Neyman-Pearson formulation. We address a more general scenario, where each hypothesis corresponds to a cluster of distributions. Such a nonparametric problem has not been thoroughly explored in the literature. We develop two nonparametric tests based respectively on the maximum mean discrepancy (MMD) and the Kolmogorov-Smirnov (KS) distance, and characterize the *exponential* error decay rate for these tests. Furthermore, in contrast to previous works where the number of hypotheses is assumed fixed, we study the regime where the number of hypotheses scales along with the sample size. This is analogous to the information theoretic channel coding problem where the number of messages scales along with the codeword length. Hence, in our study, information theory not only provides a technical tool to analyze the performance, but also provides an asymptotic perspective for understanding nonparametric hypothesis testing problems in the regime where the number of hypotheses is large, i.e., in the large-hypothesis large-sample regime.

More specifically, this paper assumes that there are M hypotheses, each corresponding to a cluster of distributions, which are *unknown*. During the training phase, sequences of length-n training data samples generated by each distribution are available. The more general case with the training data sequences having different lengths is discussed in Section III-D. Then, during the testing phase, a length n data stream is observed, consisting of samples generated by one of the distributions. The goal is to determine the cluster that contains the distribution that generated the observed test sequence. We are interested in the large-hypothesis regime, in which $M = 2^{nD}$, i.e., the number of hypotheses scales exponentially in the number of samples with a constant rate D. The analogy to the channel coding problem [1] is now apparent where the exponent represents the transmission rate, i.e., the transmitted bits per channel use, for the channel coding problem, here D represents the number of *hypothesis bits* that can be distinguished per observation

Manuscript received December 8, 2017; revised April 17, 2018 and July 27, 2018; accepted August 1, 2018. Date of publication August 17, 2018; date of current version September 27, 2018. This work was supported in part by the Defense Advanced Research Projects Agency under Contract HR0011-16-C-0135. The work of Y. Liang was supported by the National Science Foundation under Grant CCF-1801855. The work of B. Chen was supported in part by the DDDAS program of the Air Force Office of Scientific Research under Grant FA9550-16-1-0077 and in part by the National Science Foundation under Grant CNS-1731237. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Miguel Rodrigues. *(Corresponding author: Qunwei Li.)*

Q. Li, T. Wang, B. Chen, and P. K. Varshney are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (e-mail: qli33@syr.edu; twang17@syr.edu; bichen@syr.edu; varshney@syr.edu).

D. J. Bucci is with Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ 08002 USA (e-mail: Donald.J.Bucci.Jr@lmco.com).

Y. Liang is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: liang.889@osu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2018.2865884

sample. Correspondingly, we refer to D as the *discrimination rate*, and the largest such value is referred to as the *discrimination capacity*. The notion of *discrimination capacity* provides the fundamental performance limit for the hypothesis testing problem in the large-hypothesis and large-sample regime.

A. Main Contributions

This paper makes the following major contributions.

- We provide an asymptotic viewpoint to understand the nonparametric hypothesis testing problem in the regime where the number of hypotheses scales exponentially in the sample size. Based on its connection to the channel coding problem, we introduce the notions of the discrimination rate and the discrimination capacity as the performance metrics in such an asymptotic regime.
- We develop two nonparametric approaches that are based respectively on the maximum mean discrepancy (MMD) and the Kolmogorov-Smirnov (KS) distance. For both tests, we derive the error exponents and the discrimination rates. Our results show that as long as the number M of hypotheses does not scale too fast, i.e., the scaling (discrimination) exponent is less than a certain threshold, the derived tests are exponentially consistent. For each algorithm, the proof of its discrimination rate is similar to the *achievability* proof in channel coding.
- We also derive an upper bound on the discrimination capacity, which serves as an upper limit beyond which exponential consistency cannot be achieved by any nonparametric composite hypothesis testing rule. This upper bound is based on the Fano minimax method, and is similar to the *converse* proof used in channel coding.

B. Related Work

Parametric hypothesis testing: For *parametric* hypothesis testing problems, information theoretic tools have been developed to characterize the error exponent [1]–[3], [8], and to study a class of distributed parametric hypothesis testing problems [4], [5], [9]. For sequential multi-hypothesis testing, information theoretic bounds on the sample size subject to constraints on the error probabilities have been developed in [10]. A generalization of the classical hypothesis testing problem is studied in [11], where a Bayesian decision maker is designed to enhance its information about the correct hypothesis. Information theory has also been applied to study *nonparametric* hypothesis testing problems with the primary focus being on the Neyman-Pearson formulation [6], [7]. An information-theoretic approach to the problem of a nonparametric hypothesis test with a Bayesian formulation is presented in [12]. By factorizing dependent variables into mutually independent subsets, it has been shown that the likelihood ratio can be written as the sum of two sets of Kullback-Leibler divergence (KLD) terms, which is then used to quantify loss in hypothesis separability. Our study is different in that we focus on the asymptotic regime where the number of hypotheses scales with sample size.

Supervised learning: The problem we study here can also be viewed as a supervised learning problem studied in the machine learning literature. However, the problem formulated here is different from the traditional supervised learning problem [13], where sample points corresponding to the same label are treated as individual samples, and their underlying statistical structure is not exploited in the design of classification rules. For example, the support vector machine (SVM) is one of the important classification algorithms for supervised learning, where the distance between samples is measured either by the Euclidean distance or by a kernel-based distance. Such distances do not exploit the underlying statistical distributions of data samples. A robust form of the SVM in [14] incorporates the probabilistic uncertainty into the maximization of the margin. Our formulation exploits the underlying probabilistic structure of data samples, which is also robust to missing data, system noise, etc.

A formulation of the supervised learning problem that is similar to our formulation has been studied previously in [15]. The proposed approach, therein named support measure machine (SMM), exploits the kernel mean embedding to estimate the distance between probability distributions. In fact, the comparison between an SMM and an SVM also reflects the differences between our formulation and the traditional supervised learning problem. However, the study in [15] focused only on the regime with finite and fixed number of classes, and did not characterize the decay exponent of the error probability, whereas our focus is mainly on the asymptotic regime with infinite number of classes, and on the scaling behavior of the number of classes under which an asymptotically small error probability can be guaranteed. Nevertheless, the kernel-based approach developed in [15] as well as in various other papers [16]–[18] provide important techniques that we exploit in our study.

Information theory in learning: Quite a few recent studies have applied various notions in information theory for studying supervised learning problems. A minimax approach for supervised learning, where the goal is to minimize the worst-case expected loss function over a certain set of probability distributions was developed in [19]. The designed classification rules are expected to be robust over datasets generated by any probability distribution in the set. A classification problem, where the observation is obtained via a linear mapping of a vector input was studied in [20]. The notion of classification capacity was proposed, which is similar to the discrimination capacity we propose. However, the results in [20] are derived under the Gaussian model, whereas our formulation does not assume any specific distributions and is hence much more general. Furthermore, a parametric setting is implicitly assumed in [20], whereas our focus is on the nonparametric problem. A connection between the hypothesis testing problems and channel coding was established in [20], whereas in this paper we focus on the asymptotic case where the number of classes can grow exponentially large. A supervised learning problem, where the joint distribution of the data sample and its label is assumed to be known but with an unknown parameter, was studied in [21]. A classifier was proposed and the corresponding performance was analyzed. The connection of the problem to rate-distortion theory was explored. There are several key differences between the work in [21] and our study. There is no notion of discrimination rate in [21], and the performance is not defined in terms

of the asymptotic classification error probability. Additionally, our study does not assume any joint distribution of both the data sample and its label.

II. Problem Formulation

In this section, we first describe our composite nonparametric hypothesis model, and then connect it to the channel coding problem, which motivates several information theory related definitions that we will use to characterize system performance. For ease of readability, we also give preliminaries on the parametric hypothesis testing problem.

A. Supervised Learning as Nonparametric Hypothesis Testing

Consider the following nonparametric hypothesis testing problem with composite distributions. Suppose there are M hypotheses, and each hypothesis corresponds to a set $\mathcal{P}_m$ of distributions for $m = 1, \ldots, M$. For a given distance measure $d(p, q)$ between two probability distributions p and q, define

$$d(\mathcal{P}_m) := \sup_{p_i, p_{i'} \in \mathcal{P}_m} d(p_i, p_{i'}),$$

$$d(\mathcal{P}_m, \mathcal{P}_{m'}) := \inf_{p_i \in \mathcal{P}_m, p_{i'} \in \mathcal{P}_{m'}} d(p_i, p_{i'}) \quad \text{for } m \neq m'. \tag{1}$$

Hence, $d(\mathcal{P}_m)$ represents the diameter of the m-th distribution set and $d(\mathcal{P}_m, \mathcal{P}_{m'})$ represents the inter-set distance between the mth and the m'th sets.

We assume that

$$\limsup_{M \to \infty} \sup_{m=1,\ldots,M} d(\mathcal{P}_m) < D_I,$$

$$\liminf_{M \to \infty} \inf_{\substack{m, m'=1,\ldots,M \\ m \neq m'}} d(\mathcal{P}_m, \mathcal{P}_{m'}) > D_O, \tag{2}$$

where $D_I < D_O$. That is, the intra-set distance (diameter) is always smaller than the inter-set distance for the composite hypothesis testing problem. The actual values of D_I and D_O depend on the distance metrics used. Furthermore, $\limsup_{M \to \infty}$ and $\liminf_{M \to \infty}$ in (2) require that the conditions hold in the limit of asymptotically large M, i.e., the limit taken over the sequences of distribution clusters. We study the case where none of the distributions in the sets $\mathcal{P}_m$ for $m = 1, \ldots, M$ are known. Instead, for $m = 1, \ldots, M$, we assume that each distribution $p_{m,i_m} \in \mathcal{P}_m$, where $i_m \in I_1^{M_m} = \{1, 2, \ldots, M_m\}$ is the index of the distribution, generates one training sequence $\mathbf{x}_{m,i_m} \in \mathbb{R}^n$ consisting of n independently and identically distributed (i.i.d.) scalar training samples. We use $\mathbf{X}_m$ to denote all training sequences generated by the distributions in $\mathcal{P}_m$. We assume that a test sequence $\mathbf{y} \in \mathbb{R}^n$ of n i.i.d. scalar samples is generated by one of the distributions in one of the sets $\mathcal{P}_m$. The goal is to determine the hypothesis that the test sequence $\mathbf{y}$ belongs to, i.e., which set contains the distribution that generated $\mathbf{y}$.

A practical example of the considered problem involves nonparametric detection of micro-Doppler modulated radar returns, such as those which occur in a ground moving target indicator (GMTI) radar [22]. The micro-Doppler motion of a particular target generates a specific sideband structure, which varies within a distributional radius as the fundamental frequency of the target's micro motion changes, i.e., D_I. The difference between the fundamental sideband structure of the micro-Doppler modulations for different target types implies a distributional difference, i.e., D_O. This problem is clearly composite (based on an unknown fundamental modulation frequency), and a parametric realization is in many cases impractical as the specific physics of the movement can be very difficult to model in a closed form.

Let $\delta(\{\mathbf{X}_m\}_{m=1}^M, \mathbf{y})$ denote a test based on the given data. Then, the error probability for δ is defined as

$$P_e = \sum_{m_0=1}^{M} P\left(\delta(\{\mathbf{X}_m\}_{m=1}^M, \mathbf{y}) \neq m_0 \,\middle|\, \mathbf{y} \sim p_{m_0,j} \in \mathcal{P}_{m_0}\right) \cdot P(m_0), \tag{3}$$

where $P(m_0)$ is the *a priori* probability that $\mathbf{y}$ is drawn from the m_0-th set of distributions.

For the above M-ary hypothesis testing problem, we are interested in the regime, in which the number M of hypotheses scales with the number of samples. In particular, we assume $M = 2^{nD}$, where the parameter D captures how fast M scales with n. We refer to D as the *discrimination rate*.

Definition 1: We say that the discrimination rate D is achievable, if there exists a classification rule δ such that the probability of error converges to zero as the number n of observation samples converges to infinity.

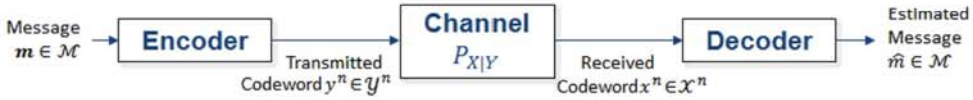
For a given composite hypothesis testing problem, we define the largest possible discrimination rate, D, to be the *discrimination capacity*, and denote it as $\bar{D}$.

B. Connection to the Channel Coding Problem

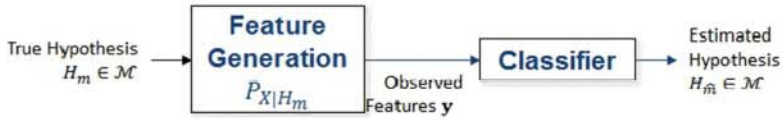
Next, we discuss the connection between the asymptotic regime of the hypothesis testing problem and the channel coding problem studied in communications, which in fact motivated our definition of the discrimination rate and the discrimination capacity.

In the channel coding problem (see Fig. 1(a)), assume there are $\mathcal{M} = \{1, \ldots, 2^{nR}\}$ messages to be transmitted with equal probability. An encoder maps each message $m \in \mathcal{M}$ one-to-one onto a length-n codeword $y_m^n = \{y_{m1}, \ldots, y_{mn}\}$, which is transmitted over the channel. The channel maps each input symbol to an output symbol in a discrete memoryless fashion with the transition probability $P_{X|Y}(x|y)$ for each channel use, and the corresponding output sequence is given by $x^n = \{x_1, \ldots, x_n\}$. A decoder then estimates the original message as $\hat{m}$ based on the output sequence. Essentially, in the channel coding problem, there are a total of M possible conditional distributions $p_m(x^n) = P_{X|Y}(x^n|y_m^n)$ given y_m^n, where $m = 1, \ldots, M$, and the decoder determines which distribution $p^* \in \{p_1, \ldots, p_M\}$ most probably generated the observed channel output x^n.

The decoding process of the channel coding problem described above is a hypothesis testing problem. Inspired by the channel coding problem, our total number of hypotheses corresponds to the total number of messages in channel coding,



(a) An illustration of the channel coding problem.



(b) An illustration of the multiple hypothesis testing problem.

Fig. 1. Illustrations of the channel coding problem and the multiple hypothesis testing problem.

and the discrimination rate D we define corresponds to the communication rate R in channel coding, which represents the transmitted message bits per coded symbol. By analogy, the discrimination rate D can be interpreted as the number of class-bits that can be distinguished per observation sample. Similarly, the discrimination capacity $\bar{D}$ corresponds to the capacity in channel coding, and serves as the fundamental limit in hypothesis testing problems. Note that in channel coding, the transmitter can choose to shape the distributions of transmitted symbols. Here, the hypothesis testing problem corresponds to the case where the distributions remain unshaped.

Essentially, Shannon's channel coding theorem guarantees error-free transmission of an exponentially increasing number of messages provided that the transmission rate R is less than the channel capacity C. In other words, Shannon's theorem implies that codewords $\{y^n\}$ can be designed such that exponentially increasing number of conditional probability distributions can be distinguished given the channel output. Here, for the hypothesis testing problem, channel coding motivates us to investigate the following problems:

- Which tests distinguish an exponentially increasing number of hypotheses with asymptotically small error probability based on n observation samples?
- What are the corresponding discrimination rates?

C. Preliminaries on Parametric Hypothesis Testing

The aforementioned questions can be answered for the *parametric* hypothesis testing problem in the asymptotic regime based on existing studies, e.g., [1]. We explain this in detail for single distributions below as preliminary material before we delve into the main focus of this paper on the *nonparametric composite hypothesis testing problem*.

Consider the *parametric* hypothesis testing problem, where there are $M = 2^{nD}$ *known* distinct distributions $p_1, ..., p_M$ corresponding respectively to M hypotheses. Given a test sequence $\mathbf{y}$ consisting of n i.i.d. samples generated from one of these distributions, the goal is to determine the true hypothesis, i.e., which distribution p_i generated the test sequence.

We apply the likelihood test given by:

$$\delta(\mathbf{y}) = \arg\max_i P_{X|H_i}(\mathbf{y}) \tag{4}$$

where the test labels the observed test data as hypothesis i if p_i generates $\mathbf{y}$ with the largest probability. It can be shown [1] that the likelihood test in (4) is equivalently given by

$$\delta(\mathbf{y}) = \arg\min_i D_{KL}(\gamma(\mathbf{y})\|p_i), \tag{5}$$

where $D_{KL}(\cdot\|\cdot)$ is the KLD between two distributions, and $\gamma(\cdot)$ is the empirical distribution of the sequence. It suggests that the testing rule labels the test data as hypothesis i if the empirical distribution of the test data is closest to p_i in KLD.

We next analyze the average error probability of the above testing rule as follows.

$$\begin{aligned} P_e &= \frac{1}{M}\sum_{j=1}^{2^{nD}} P(\delta(\mathbf{y}) \neq j|H_j) = \frac{1}{M}\sum_{j=1}^{2^{nD}} P(\exists\, i \neq j \text{ s.t. } \mathcal{E}_1|H_j) \\ &\leq \frac{1}{M}\sum_{j=1}^{2^{nD}}\sum_{i,i\neq j} P(\mathcal{E}_1|H_j) = \frac{1}{M}\sum_{j=1}^{2^{nD}}\sum_{i,i\neq j} \exp\{-nC(p_i,p_j)\} \\ &\leq 2^{nD - n\log e \liminf_{M\to\infty} \min_{1\leq i,j\leq M} C(p_i,p_j)}, \end{aligned} \tag{6}$$

where $C(p_i, p_j)$ denotes the Chernoff distance

$$C(p_i,p_j) = \max_{0\leq t\leq 1} -\log \int [p_i(p)]^{1-t}[p_j(p)]^t dp, \tag{7}$$

and $\mathcal{E}_1$ denotes the event that given H_j, the KLD between $\mathbf{y}$ and p_j is greater than the KLD between $\mathbf{y}$ and p_i for some $i \neq j$, i.e., for $i \neq j$, $D_{KL}(\gamma(\mathbf{y})\|p_i) < D_{KL}(\gamma(\mathbf{y})\|p_j)$. Note that for simplicity, the default base for log in this paper is 2. Thus, if $D \leq \log e \liminf_{M\to\infty} \min_{1\leq i,j\leq M} C(p_i,p_j)$, then the error probability is asymptotically small as n goes to infinity, which proves the following proposition.

Proposition 1: For the parametric multiple hypothesis testing problem, the discrimination rate D is achievable if

$$D \leq \log e \liminf_{M\to\infty} \min_{1\leq i,j\leq M} C(p_i,p_j).$$

Hence, for the discrimination rate to be positive, we require that the smallest pairwise Chernoff information be bounded away from zero for asymptotically large M, i.e., the limit taken over the sequences of distribution clusters.

III. Main Results

In this section, we obtain the performance bounds for the nonparametric hypothesis testing problem, with two different distance measures, i.e., MMD and KS distance.

A. MMD-Based Test

We construct a nonparametric test based on the MMD between two distributions p and q [18] defined as

$$\mathrm{MMD}^2(p,q) := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{8}$$

where $\mu_p(\cdot)$ maps a distribution p into an element in a reproducing kernel Hilbert space (RKHS) associated with a kernel $k(\cdot,\cdot)$ as

$$\mu_p(\cdot) = \mathbb{E}_p[k(\cdot,x)] = \int k(\cdot,x)dp(x). \tag{9}$$

An unbiased estimator of (8) based on n samples of $\mathrm{x} = \{x_1,\ldots,x_n\}$ generated by distribution p and m samples of $\mathrm{y} = \{y_1,\ldots,y_m\}$ generated by distribution q, is [18]:

$$\mathrm{MMD}^2(\mathrm{x},\mathrm{y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(y_i,y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j). \tag{10}$$

Note that $x_i, y_i \in \mathbb{R}^d$, and the dimension $d \geq 1$.

We employ the MMD to measure the distance between the test sequence and the training sequences, and declare the hypothesis of the test sequence to be the same as the training sequence that has the smallest MMD to the test sequence. The constructed MMD-based nonparametric composite hypothesis test is given by

$$\delta_{\mathrm{MMD}}(\{\mathbf{X}_m\}_{m=1}^{M},\mathrm{y}) = \arg\min_{m,i_m} \mathrm{MMD}^2(\mathrm{x}_{m,i_m},\mathrm{y}). \tag{11}$$

The following theorem characterizes the average probability of error performance of the proposed MMD-based test under composite distributions.

Theorem 1: Suppose the MMD-based test in (11) is applied to the nonparametric composite hypothesis testing problem under assumption (2), where the kernel satisfies $0 \leq k(x,y) \leq \mathcal{K}$ for all (x,y). Then, the average probability of error is upper bounded as

$$P_e \leq 2^{nD}\exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right).$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{96\mathcal{K}^2}(D_O - D_I)^2. \tag{12}$$

Proof: See Appendix A. ■

Next, we study a special case where each hypothesis is associated with a single distribution, i.e., the m-th hypothesis is associated with only one distribution $p_m, m = 1,\ldots,M$. Then, we have the following corollary.

Corollary 1: Suppose the MMD-based test is applied to the nonparametric hypothesis testing problem under assumption (2), and each hypothesis is associated with a single distribution, where the kernel satisfies $0 \leq k(x,y) \leq \mathcal{K}$ for all (x,y). Then, the average probability of error under equally probable hypotheses is upper bounded as

$$P_e \leq 2^{nD - n\frac{\log e}{96\mathcal{K}^2}\liminf_{M\to\infty}\min_{1\leq i,j\leq M}\mathrm{MMD}^4(p_i,p_j)}. \tag{13}$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{96\mathcal{K}^2}\liminf_{M\to\infty}\min_{1\leq i,j\leq M}\mathrm{MMD}^4(p_i,p_j). \tag{14}$$

Note that, for the discrimination rate to be positive, we require the smallest pairwise MMD between the distributions to be bounded away from zero for asymptotically large M, where the limit is taken over the sequences of distribution clusters.

Proof: By Theorem 1, we set $D_I = 0$ and $D_O = \liminf_{M\to\infty}\min_{1\leq i,j\leq M}\mathrm{MMD}^2(p_i,p_j)$. Therefore, we can bound the probability of error as the number of classes scales according to $M = 2^{nD}$

$$P_e \leq M\exp\left(-\frac{n\liminf_{M\to\infty}\min_{1\leq i,j\leq M}\mathrm{MMD}^4(p_i,p_j)}{96\mathcal{K}^2}\right) \leq 2^{nD - n\frac{\log e}{96\mathcal{K}^2}\liminf_{M\to\infty}\min_{1\leq i,j\leq M}\mathrm{MMD}^4(p_i,p_j)}. \tag{15}$$

Then, it is straightforward to obtain the achievable discrimination rate for the MMD test as

$$D = \frac{\log e}{96\mathcal{K}^2}\liminf_{M\to\infty}\min_{1\leq i,j\leq M}\mathrm{MMD}^4(p_i,p_j). \tag{16}$$

■

B. Kolmogorov-Smirnov Test

In this section, we construct a nonparametric hypothesis testing test based on the KS distance defined as follows. Suppose $\mathrm{x} = \{x_1,\ldots,x_n\}$, and i.i.d. samples $x_i \in \mathbb{R}$, are generated by the distribution p. Then the empirical CDF of p is given by

$$F_{\mathrm{x}}(a) = \frac{1}{n}\sum_{i=1}^{n} 1_{[-\infty,a]}(x_i), \tag{17}$$

where $1_{[-\infty,x]}$ is the indicator function. The KS distance between x and y having respectively been generated by p and q is defined as

$$D_{KS}(\mathrm{x},\mathrm{y}) = \sup_{a\in\mathbb{R}}|F_{\mathrm{x}}(a) - F_{\mathrm{y}}(a)|. \tag{18}$$

We construct the following KS based nonparametric composite hypothesis test

$$\delta_{KS}(\{\mathbf{X}_m\}_{m=1}^{M},\mathrm{y}) = \arg\min_{m,i_m} D_{KS}(\mathrm{x}_{m,i_m},\mathrm{y}), \tag{19}$$

The following theorem characterizes the performance of the proposed KS-based test.

Theorem 2: Suppose the KS-based test in (19) is applied to the nonparametric hypothesis testing problem under assumption (2). Then, the average probability of error is upper bounded as

$$P_e \leq 6\cdot 2^{nD}\exp\left(-\frac{n(D_O - D_I)^2}{8}\right).$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{8}(D_O - D_I)^2. \tag{20}$$

Proof: See Appendix B. ■

Consider the case where each hypothesis is associated with a single distribution, and we have the following corollary.

Corollary 2: Suppose the KS-based test is applied to the nonparametric hypothesis testing problem under assumption (2), and each hypothesis is associated with a single distribution. Then, the average probability of error under equally probable hypotheses is upper bounded as

$$P_e \leq 6 \cdot 2^{nD - n\frac{\log e}{8} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} d_{KS}^2(p_i,p_j)}. \tag{21}$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{8} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} d_{KS}^2(p_i,p_j). \tag{22}$$

Hence, for the discrimination rate to be positive, we require the least pairwise KS distance between distributions to be bounded away from zero for asymptotically large M, where the limit is taken over the sequences of distribution clusters.

Proof: By Theorem 2, we set $D_I = 0$ and $D_O = \liminf_{M\to\infty} \min_{1\leq i,j\leq M} d_{KS}(p_i,p_j)$, and have

$$P_e \leq 6 \cdot 2^{nD - n\frac{\log e}{8} D_o^2} \leq 6 \cdot 2^{nD - n\frac{\log e}{8} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} d_{KS}^2(p_i,p_j)}.$$

Then, it is straightforward to obtain the achievable discrimination rate for the KS test. ■

C. Upper Bound on the Discrimination Capacity

In this section, we provide an upper bound on the discrimination capacity for the composite hypothesis testing problem. Let h be a random index representing the actual hypothesis that occurs. We assume that h is uniformly distributed over the M hypotheses, and h' has the same distribution as m, but is independent from h. Then, Lemma 2.10 in [23] directly yields the following upper bound on the discrimination capacity $\bar{D}$.

Remark 1: The discrimination capacity $\bar{D}$ is upper bounded as

$$\bar{D} \leq \limsup_{M\to\infty} \mathbb{E}_{h,h'} D_{KL}(p_h \| p_{h'}), \tag{23}$$

where $D_{KL}(\cdot\|\cdot)$ is the KLD between two distributions.

Note that the above limit $\limsup_{M\to\infty}$ is taken over the sequences of distribution clusters. In Appendix C, we provide an alternative but simpler proof based on Fano's inequality for the above upper bound, which is closely related to the proposed concept of discrimination capacity.

D. Training Sequences of Unequal Length

In this subsection, we discuss the impact of different number of training samples in different classes on the probability of error and the discrimination rate. Here, we still assume that there are n test samples. To keep the problem formulation meaningful, we assume that the number M of classes increases exponentially with n at a rate D, i.e., $M = 2^{nD}$. To avoid notational confusion, we use the non-composite case, i.e., with each class corresponding to one distribution, to illustrate the idea. Suppose that each class, i.e., each distribution, generates $\gamma_m(n)$ training samples, for $m = 1, \ldots, M$, where $\gamma_m(n)$ represents the number of samples in the m-th class (as a function of n). Let $\gamma_{\min}(n) = \min_{1\leq m\leq M} \gamma_m(n)$. In particular, for the MMD-based test, the probability of error can be bounded as

$$P_e \leq 2^{n\left(D - \min\{1, \frac{\gamma_{\min}(n)}{n}\} \frac{\log e (D_O - D_I)^2}{96K^2}\right)}. \tag{24}$$

For the KS-based test, the probability of error can be bounded as

$$P_e \leq 6 \cdot 2^{n\left(D - \min\{1, \frac{\gamma_{\min}(n)}{n}\} \frac{\log e (D_O - D_I)^2}{8}\right)}. \tag{25}$$

It can be seen that here the ratio $\frac{\gamma_{\min}(n)}{n}$ plays an important role in determining the error exponent asymptotically. For example, for the MMD-based test, if the ratio converges to zero for large n, i.e., the shortest training length $\gamma_{\min}(n)$ scales as an order-level slower than the test length, then there is no guarantee of exponential error decay, and the discrimination rate equals zero. On the other hand, if $\lim_{n\to\infty} \frac{\gamma_{\min}(n)}{n} = c$ with $0 < c < 1$, then the discrimination rate $D = c\frac{\log e (D_O - D_I)^2}{96K^2}$. Furthermore, if $\lim_{n\to\infty} \frac{\gamma_{\min}(n)}{n} = c$ with $c \geq 1$, then the discrimination rate $D = \frac{\log e (D_O - D_I)^2}{96K^2}$. A sketch of the proof of (24) and (25) can be found in Appendix D

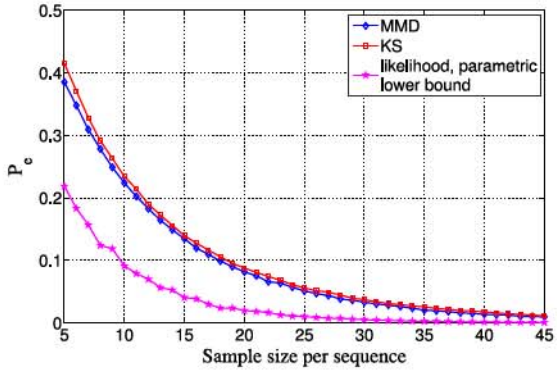


Fig. 2. Error probabilities of different hypothesis testing algorithms for Gaussian distributions with different means.

IV. Numerical Results

In this section, we present numerical results to compare the performance of the proposed tests. In the experiment, the number of classes is set to be five, and the error probability versus the number of samples for the proposed algorithms is plotted. For the MMD based test, we use the standard Gaussian kernel given by $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2})$.

In the first experiment, all the hypotheses correspond to Gaussian distributions with the same variance $\sigma^2 = 1$ but different mean values $\mu = \{-2, -1, 0, 1, 2\}$. A training sequence is drawn from each distribution and a test sequence is randomly generated from one of the five distributions. The sample size of each sequence ranges from 5 to 45. A total of 10^5 monte carlo runs are conducted. The simulation results are given in Fig. 2. It can be seen that all the tests give better performance as the sam-

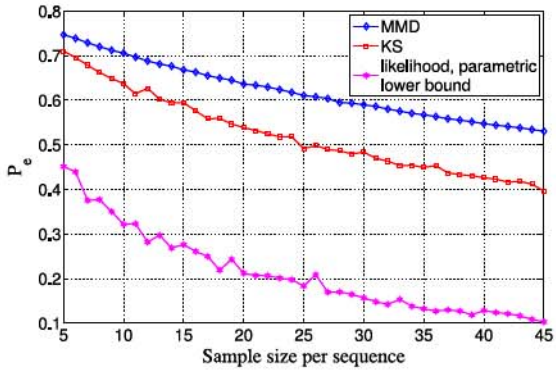


Fig. 3. Error probabilities of different hypothesis testing algorithms for Gaussian distributions with different variances.

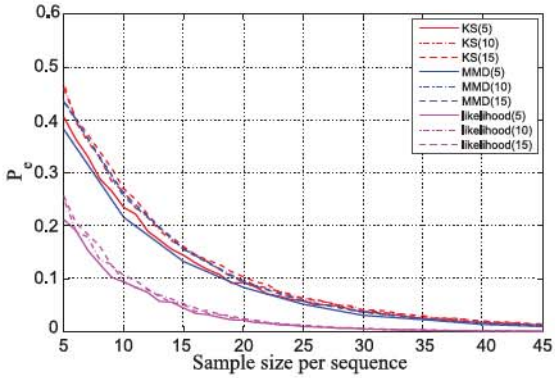


Fig. 4. Error probabilities for different hypothesis testing algorithms for Gaussian distributions with different means.

TABLE I
COMPARISON OF BOUNDS

	Lower Bounds		Upper Bounds	
	KS	MMD	Parametric	FB
Empirical	0.0897	0.0916	0.146	2.5
Theoretical	0.0183	0.0071	0.125	-

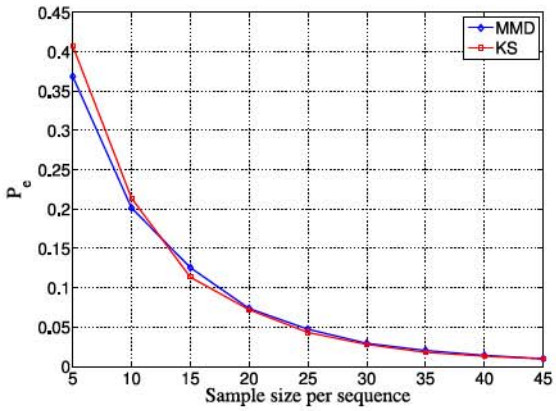


Fig. 5. Error probabilities of different hypothesis testing algorithms for composite Gaussian distributions with different means.

ple size n increases. We can also see that the MMD-based test slightly outperforms the KS-based test. We also provide results for the parametric likelihood test as a lower bound on the probability of error for performance comparison. It can be seen that the performance of the two nonparametric tests are close to the parametric likelihood test even with a moderate number of samples.

In the second experiment, all the hypotheses correspond to Gaussian distributions with the same mean $\mu = 1$ but different variance values $\sigma^2 = \{0.5^2, 1^2, 1.5^2, 2^2, 2.5^2\}$. The simulation results are given in Fig. 3. In this experiment, the MMD-based test yields the worst performance, which suggests that this method is not suitable when the distributions overlap substantially with each other. The two simulation results also suggest that none of the three tests perform the best universally over all distributions. Although there is a gap between the performance of MMD and KS tests and that of the parametric likelihood test, we observe that the error decay rates of these tests are still close.

To show the tightness of the bounds derived in the paper, we provide a table (See Table I) of error decay exponents (and thus the discrimination rates) for different algorithms. Estimates of error decay exponent of KS and MMD based tests on a multi-hypothesis testing problem are presented for the problem considered in the first experiment. Note that the theoretical lower bounds in the table correspond to the achievable discrimination rates of the methods asymptotically. Fano's bound (FB in the table) is estimated by using data-dependent partition estimators of Kullback-Leibler divergence [24]. The parametric upper bound is based on the maximum likelihood test, which can serve as an upper bound on the error decay exponent (and hence intuitively on the discrimination capacity). It can be seen from the table that both the KS and MMD tests do achieve an exponential error decay and have positive discrimination rates as we show in our theorems. Clearly, the empirical values of the bounds for both tests are better than the corresponding theoretical values. More importantly, both of the empirical lower bounds are close to the likelihood upper bound, demonstrating that the actual performance of the two tests are satisfactory. We also note that the Fano's upper bound is not very close to the lower bound.

To better illustrate the bounds in Table I, we provide experimental results with different number of hypotheses M in Fig. 4. In particular, we present the simulation results with $M = 5, 10, 15$. We use a similar experiment setting as that in the first experiment, where Gaussian distributions have the same variance and different mean values, and the mean values are $\{-2, -1, \ldots, 2\}, \{-4.5, -3.5, \ldots, 4.5\}$ and $\{-7, -6, \ldots, 7\}$ respectively. The parametric maximum likelihood test serves as an unpper bound for the error decay exponent for all of the three cases. Similar to the case $M = 5$, KS and MMD nonparametric tests achieve an exponential error decay and hence the positive discrimination rates for the cases $M = 10$ and $M = 15$.

We now conduct experiments with composite distributions. First, we still use five hypotheses with Gaussian distributions with variance $\sigma^2 = 1$ and different mean values $\mu = \{-2, -1, 0, 1, 2\}$. For each hypothesis, we vary the mean values by ± 0.1. Thus, within each hypothesis, there are three different distributions with mean values in $\{\mu - 0.1, \mu, \mu + 0.1\}$. The results are presented in Fig. 5. As expected, the performance improves as the sample size n increases. The two tests perform

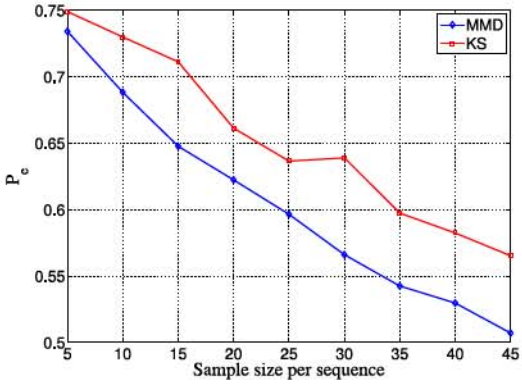


Fig. 6. Error probabilities of different hypothesis testing algorithms for composite Gaussian distributions with different variances.

almost identically, with the MMD-based test slightly outperforming the KS-based test for small n.

We again vary the variances of the Gaussian distributions as in the second experiment in a similar way. In particular, the variances in the same class are $\{(\sigma-0.1)^2, \sigma^2, (\sigma+0.1)^2\}$, and $\sigma \in \{0.5, 1, 1.5, 2, 2.5\}$. In Fig. 6, we observe the performance improvement as the sample size n increases. Different from the results in the second experiment, the MMD-based test outperforms the KS-based test in the composite setting.

V. Conclusion

This paper developed a nonparametric composite hypothesis testing approach for arbitrary distributions based on the maximum mean discrepancy (MMD) and Kolmogorov-Smirnov (KS) distance measure based tests. We introduced the information theoretic notion of discrimination capacity that was defined for the regime where the number of hypotheses scales along with the sample size. We also provided characterization of the corresponding error exponent and the discrimination rate, i.e., a lower bound on the discrimination capacity. Our framework can been extended to unsupervised learning problems and similar performance limits can be investigated.

Appendix

A. Proof of Theorem 1

The proof uses the following inequality.

Lemma 1: [McDiarmid's Inequality [25]] Let $f : \mathcal{X}^m \to \mathbb{R}$ be a function such that for all $i \in \{1, \ldots, m\}$, there exist $c_i \leq \infty$ for which

$$\sup_{X \in \mathcal{X}^m, \tilde{x} \in \mathcal{X}} |g(x_1, \ldots, x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)| \leq c_i, \quad (26)$$

where $g(x_1, \ldots, x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m) = f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)$. Then for all probability measure p and every $\epsilon > 0$,

$$P_X\left(f(X) - \mathbb{E}_X[f(X)] > \epsilon\right) < \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right), \quad (27)$$

where X denotes $(x_1, \ldots, x_m)$, $\mathbb{E}_X[\cdot]$ denotes the expectation over the m random variables $x_i \sim p$, and P_X denotes the probability over these m variables.

To apply the McDiarmid's inequality, we first define the following quantity

$$\triangle_m(\mathbf{x}^\alpha) = \mathrm{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) - \mathrm{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y}) \quad (28)$$

where $\mathbf{x}^\alpha := \{\mathbf{x}_{m,i_m}, \mathbf{x}_{m',i_{m'}}, \mathbf{y}\}$ consists of $3n$ data samples.

Given H_i, it can be shown that

$$\mathbb{E}[\mathrm{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y})] \leq D_I \quad (29)$$

$$\mathbb{E}[\mathrm{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y})] \geq D_O. \quad (30)$$

We next define $\mathbf{x}^\alpha_{-s}$ the same as $\mathbf{x}^\alpha$ except that the s-th component $\mathbf{x}^\alpha_s$ is removed. We also define $\tilde{\mathbf{x}}^\alpha_s$ as another sequence generated by the same underlying distribution for $\mathbf{x}^\alpha_s$. Then, $\mathbf{x}^\alpha_s$ affects $\triangle_m(\mathbf{x}^\alpha)$ via the following three cases.

- Case 1: $\mathbf{x}^\alpha_s$ is in the sequence $\mathbf{x}_{m,i_m}$. In this case, $\mathbf{x}^\alpha_s$ affects $\triangle_m(\mathbf{x}^\alpha)$ through the following terms
$$\frac{2}{n(n-1)} \sum_{l=1, l\neq s}^{n} k(\mathbf{x}^\alpha_s, \mathbf{x}_{m,i_m}(l)) - \frac{2}{n^2} \sum_{l=1}^{n} k(\mathbf{x}^\alpha_s, \mathbf{y}(l)).$$
- Case 2: $\mathbf{x}^\alpha_s$ is in the sequence $\mathbf{x}_{m',i_{m'}}$. In this case, $\mathbf{x}^\alpha_s$ affects $\triangle_m(\mathbf{x}^\alpha)$ through the following terms
$$\frac{2}{n(n-1)} \sum_{l=1, l\neq s}^{n} k(\mathbf{x}^\alpha_s, \mathbf{y}(l)) - \frac{2}{n^2} \sum_{j=1}^{n} k(\mathbf{x}^\alpha_s, \mathbf{x}_{m',i_{m'}}(l)).$$
- Case 3: $\mathbf{x}^\alpha_s$ is in the sequence $\mathbf{y}$. In this case, $\mathbf{x}^\alpha_s$ affects $\triangle_m(\mathbf{x}^\alpha)$ through the following terms
$$\frac{2}{n^2} \sum_{l=1}^{n} k(\mathbf{x}^\alpha_s, \mathbf{x}_{m,i_m}(l)) - \frac{2}{n^2} \sum_{l=1}^{n} k(\mathbf{x}^\alpha_s, \mathbf{x}_{m',i_{m'}}(l)).$$

Thus, since the kernel is bounded, i.e., $0 \leq k(x;y) \leq \mathcal{K}$ for any (x, y), considering the above three cases, the variation in the value of $\triangle_m(\mathbf{x}^\alpha)$ when $\mathbf{x}^\alpha_s$ varies is bounded by $\frac{4\mathcal{K}}{n}$. Then,

$$|\triangle_m(\mathbf{x}^\alpha_{-s}, \mathbf{x}^\alpha_s) - \triangle_m(\mathbf{x}^\alpha_{-s}, \tilde{\mathbf{x}}^\alpha_s)| \leq \frac{8\mathcal{K}}{n}. \quad (31)$$

We now apply Lemma 1 and obtain

$$\begin{aligned}
&P\left(\mathrm{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) \geq \mathrm{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y})\right) \\
&= P(\mathrm{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) - \mathrm{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y}) \geq 0) \\
&= P\left(\triangle_m(\mathbf{x}^\alpha) - \mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq -\mathbb{E}[\triangle_m(\mathbf{x}^\alpha)]\right) \\
&\leq P\left(\triangle_m(\mathbf{x}^\alpha) - \mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq D_O - D_I\right) \\
&\leq \exp\left(-\frac{2(D_O - D_I)^2}{64\mathcal{K}^2\frac{3}{n}}\right) = \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right). \quad (32)
\end{aligned}$$

The first inequality is based on the results in (29) and (30) that $-\mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq D_O - D_I$. Therefore,

$$\begin{aligned}
P_e &= P\left(\exists m' \neq m, i'_m \in I_1^{M'_m}, \triangle_m(\mathbf{x}^\alpha) \geq 0, \forall i_m \in I_1^{M_m}\right) \\
&\leq \frac{1}{M} \sum_{m=1}^{M} \sum_{m'\neq m} \min_{i_m \in I_1^{M_m}} \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right) \\
&\leq M \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right). \quad (33)
\end{aligned}$$

Thus, $D = \frac{\log e}{96\mathcal{K}^2}(D_O - D_I)^2$.

B. Proof of Theorem 2

We first introduce two lemmas to help establish the theorem.

Lemma 2: [26] Suppose $\mathbf{x}$ is generated by p and $F_{\mathbf{x}}(a)$ is the corresponding empirical c.d.f.. Then

$$P\left(\sup_{a\in\mathbb{R}}\left|F_{\mathbf{x}}(a)-F_p(a)\right|>\epsilon\right)\leq 2\exp\left(-2n\epsilon^2\right).$$

Lemma 3: Suppose two distribution clusters $\mathcal{P}_1$ and $\mathcal{P}_2$ satisfy (2). Assume that for $j=1,2$, $\mathbf{x}_j\sim p_j$ satisfying $p_j\in\mathcal{P}_j$. Then for any $\mathbf{x}_3\sim p_3$ satisfying $p_3\in\mathcal{P}_1$,

$$P\left(d_{KS}(\mathbf{x}_1,\mathbf{x}_3)\geq d_{KS}(\mathbf{x}_2,\mathbf{x}_3)\right)\leq 6\exp\left(-\frac{n(D_O-D_I)^2}{8}\right).$$

Proof: By the triangle inequality and the property of supremum, we have

$$d_{KS}(\mathbf{x}_1,\mathbf{x}_3)<d_{KS}(p_1,\mathbf{x}_1)+d_1+d_{KS}(p_3,\mathbf{x}_3),$$
$$d_{KS}(\mathbf{x}_2,\mathbf{x}_3)>-d_{KS}(p_3,\mathbf{x}_3)+d_2-d_{KS}(p_2,\mathbf{x}_2).$$

where $D_I<d_1<d_2<D_O$. Then

$$\begin{aligned}&P\left(d_{KS}(\mathbf{x}_1,\mathbf{x}_3)\geq d_{KS}(\mathbf{x}_2,\mathbf{x}_3)\right)\\&\leq P\left(d_{KS}(p_1,\mathbf{x}_1)+d_{KS}(p_3,\mathbf{x}_3)+2d_{KS}(p_2,\mathbf{x}_2)>\hat{d}\right)\\&\leq P\left(d_{KS}(p_1,\mathbf{x}_1)>\frac{\hat{d}}{4}\right)+P\left(d_{KS}(p_3,\mathbf{x}_3)>\frac{\hat{d}}{4}\right)\\&+P\left(d_{KS}(p_2,\mathbf{x}_2)>\frac{\hat{d}}{4}\right)\leq 6\exp\left(-\frac{n\hat{d}^2}{8}\right).\end{aligned}$$

where $\hat{d}=d_2-d_1$. Then, we have the desired result. ■

Without loss of generality, assume that the probability that $\mathbf{y}$ is generated from p_{k,i_k} is $\frac{1}{M}$ for all $m\in\{1,\ldots,M\}$ and $i_m\in\{1,\ldots,M_m\}$. By Lemma 3 and the union bound, the probability of error is bounded by

$$\begin{aligned}P_e&\leq\sum_{m=1}^{M}\sum_{i_m=1}^{M_m}\sum_{m'\neq m}\sum_{i_{m'}=1}^{M_{m'}}P\Big(d_{KS}(\mathbf{x}_{m,i_m},\mathbf{y})\geq\\&\quad d_{KS}(\mathbf{x}_{m',i_{m'}},\mathbf{y})\big|\mathbf{y}\sim p_{m,i_m},\ i_m\in\{1,\ldots,M_m\}\big|\Big)\frac{1}{M}\\&\leq 6M\exp\left(-\frac{n(D_O-D_I)^2}{8}\right).\end{aligned}\tag{34}$$

Thus, the achievable discrimination rate is $\frac{\log e}{8}(D_O-D_I)^2$.

C. Proof of Remark 1

Here we provide an alternative proof for Remark 1, which is different from that given in [Lemma 2.10 [23]].

By Fano's inequality [1], we obtain

$$H(h|\mathbf{y})\leq 1+P_e\log(M-1).\tag{35}$$

Since h is uniformly distributed over all the hypotheses, we have

$$\begin{aligned}\log(M)=H(h)&=I(h;\mathbf{y})+H(h|\mathbf{y})\\&\leq I(h;\mathbf{y})+1+P_e\log M.\end{aligned}\tag{36}$$

Let $P_h(h)$, $P_{\mathbf{y}}(\mathbf{y})$, and $P_{h,\mathbf{y}}(h,\mathbf{y})$ represent the marginal and joint distributions of h and $\mathbf{y}$. Recall that we represent the likelihood function of $\mathbf{y}$ under m as $P(\mathbf{y}|h)=p_h(\mathbf{y})$. The mutual information between h and $\mathbf{y}$ can be expressed as

$$\begin{aligned}I(h;\mathbf{y})&=\sum_{h=1}^{M}\sum_{\mathbf{y}}P_{h,\mathbf{y}}(h,\mathbf{y})\log\frac{P_{h,\mathbf{y}}(h,\mathbf{y})}{P_h(h)P_{\mathbf{y}}(\mathbf{y})}\\&=\frac{1}{M}\sum_{h=1}^{M}\sum_{\mathbf{y}}p_h(\mathbf{y})\log\frac{p_h(\mathbf{y})}{P_{\mathbf{y}}(\mathbf{y})}\\&=\frac{1}{M}\sum_{h=1}^{M}\sum_{\mathbf{y}}p_h(\mathbf{y})\log\frac{p_h(\mathbf{y})}{\sum_{h'=1}^{M}\frac{1}{M}p_{h'}(\mathbf{y})}\\&=\frac{1}{M}\sum_{h=1}^{M}\sum_{\mathbf{y}}p_h(\mathbf{y})\left[\log p_h(\mathbf{y})-\log\sum_{h'=1}^{M}\frac{1}{M}p_{h'}(\mathbf{y})\right]\end{aligned}$$

Applying Jensen's inequality, it is further upper bounded as

$$I(h;\mathbf{y})\leq\frac{1}{M}\sum_{h=1}^{M}\sum_{\mathbf{y}}p_h(\mathbf{y})\left[\log p_h(\mathbf{y})-\sum_{h'=1}^{M}\frac{1}{M}\log p_{h'}(\mathbf{y})\right]$$

Simplifying, we finally have

$$\begin{aligned}I(h;\mathbf{y})&\leq\frac{1}{M}\sum_{h=1}^{M}\sum_{\mathbf{y}}p_h(\mathbf{y})\\&\quad\cdot\left[\sum_{h'=1}^{M}\frac{1}{M}\log p_h(\mathbf{y})-\sum_{h'=1}^{M}\frac{1}{M}\log p_{h'}(\mathbf{y})\right]\\&=\frac{1}{M}\frac{1}{M}\sum_{h=1}^{M}\sum_{h'=1}^{M}\sum_{\mathbf{y}}p_h(\mathbf{y})\log\frac{p_h(\mathbf{y})}{p_{h'}(\mathbf{y})}\\&=\frac{1}{M}\frac{1}{M}\sum_{h=1}^{M}\sum_{h'=1}^{M}nD_{KL}(p_h\|p_{h'})\\&=n\mathbb{E}_{h,h'}D_{KL}(p_h\|p_{h'}).\end{aligned}\tag{37}$$

where h' has the same distribution as h, but is independent from h. Substituting (37) into the (36), we obtain

$$\log M\leq n\mathbb{E}_{h,h'}D_{KL}(p_h\|p_{h'})+1+\log MP_e\tag{38}$$

which implies that

$$\frac{\log M}{n}\leq\frac{\mathbb{E}_{h,h'}D_{KL}(p_h\|p_{h'})}{1-P_e}+\frac{1}{n(1-P_e)}.\tag{39}$$

Since $M=2^{nD}$, we can have

$$D\leq\frac{\mathbb{E}_{m,m'}D_{KL}(p_m\|p_{m'})}{1-P_e}+\frac{1}{n(1-P_e)}.\tag{40}$$

Thus, for any test that satisfies $P_e\to 0$ as $n\to\infty$, $D\leq\limsup_{M\to\infty}E_{m,m'}D_{KL}(p_m\|p_{m'})$ as $n\to\infty$. Therefore,

$$\bar{D}\leq\limsup_{M\to\infty}\mathbb{E}_{m,m'}D_{KL}(p_m\|p_{m'}).\tag{41}$$

D. Sketch of the Proof for (24) and (25)

To prove (24), we follow the steps to obtain (32). Note that now $\mathbf{x}^\alpha := \{\mathbf{x}_{m,i_m}, \mathbf{x}_{m',i_{m'}}, \mathbf{y}\}$ consists of $n+\gamma_m(n)+\gamma_{m'}(n)$ data samples, and $|\triangle_m(\mathbf{x}^\alpha_{-s}, \mathbf{x}^\alpha_s) - \triangle_m(\mathbf{x}^\alpha_{-s}, \tilde{\mathbf{x}}^\alpha_s)| \leq \frac{8\mathcal{K}}{n'}$, where $n' \in \{n, \gamma_m(n), \gamma_{m'}(n)\}$ and the correponding choice is based on the location of $\mathbf{x}^\alpha_s$. Then, we can write

$$\begin{aligned} &P\left(\mathrm{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) \geq \mathrm{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y})\right) \\ &\leq P\left(\triangle_m(\mathbf{x}^\alpha) - \mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq D_O - D_I\right) \\ &\leq \exp\left(-\frac{2(D_O-D_I)^2}{64\mathcal{K}^2\left(\frac{1}{n}+\frac{1}{\gamma_m(n)}+\frac{1}{\gamma_{m'}(n)}\right)}\right) \\ &\leq \exp\left(-\frac{\min\{n, \gamma_{\min}(n)\}(D_O-D_I)^2}{96\mathcal{K}^2}\right). \end{aligned} \tag{42}$$

Thus, it yields

$$\begin{aligned} P_e &= P\left(\exists m' \neq m, i'_m \in I_1^{M'_m}, \triangle_m(\mathbf{x}^\alpha) \geq 0, \forall i_m \in I_1^{M_m}\right) \\ &\leq \frac{1}{M}\sum_{m=1}^{M}\sum_{m'\neq m}\min_{i_m \in I_1^{M_m}} \exp\left(-\frac{n(D_O-D_I)^2}{96\mathcal{K}^2}\right) \\ &\leq M\exp\left(-\frac{\min\{n, \gamma_{\min}(n)\}(D_O-D_I)^2}{96\mathcal{K}^2}\right). \end{aligned} \tag{43}$$

To prove (25), we follow the steps to obtain (34). Note that if the sequences $\mathbf{y}, \mathbf{x}_{m,i_m}, \mathbf{x}_{m',i_{m'}}$ have length of $n, \gamma_m(n), \gamma_{m'}(n)$ respectively, we can obtain

$$\begin{aligned} &P\left(d_{KS}(\mathbf{x}_{m,i_m}, \mathbf{y}) \geq d_{KS}(\mathbf{x}_{m',i_{m'}}, \mathbf{y})\right) \\ &\leq 2\exp\left(-\frac{n(D_O-D_I)^2}{8}\right) + 2\exp\left(-\frac{\gamma_m(n)(D_O-D_I)^2}{8}\right) \\ &+ 2\exp\left(-\frac{\gamma_{m'}(n)(D_O-D_I)^2}{8}\right) \\ &\leq 6\exp\left(-\frac{\min\{n, \gamma_{\min}(n)\}(D_O-D_I)^2}{8}\right). \end{aligned} \tag{44}$$

Thus, it yields

$$\begin{aligned} P_e &\leq \sum_{m=1}^{M}\sum_{i_m=1}^{M_m}\sum_{m'\neq m}\sum_{i_{m'}=1}^{M_{m'}} P\Big(d_{KS}(\mathbf{x}_{m,i_m}, \mathbf{y}) \geq \\ &\quad d_{KS}(\mathbf{x}_{m',i_{m'}}, \mathbf{y})\big|\mathbf{y} \sim p_{m,i_m},\ i_m \in \{1,\ldots,M_m\}\big|\Big)\frac{1}{M} \\ &\leq 6M\exp\left(-\frac{\min\{n, \gamma_{\min}(n)\}(D_O-D_I)^2}{8}\right). \end{aligned} \tag{45}$$

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[2] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó: Budapest, 1981.

[4] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987.

[5] T. S. Han and S. I. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.

[6] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401–408, Feb. 1989.

[7] E. Levitan and N. Merhav, "A competitive Neyman–Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2215–2229, Aug. 2002.

[8] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 299–320, Jan. 2008.

[9] G. Vazquez-Vilar, A. T. Campo, A. G. i Fàbregas, and A. Martinez, "Bayesian m-ary hypothesis testing: The meta-converse and Verdú-Han bounds are tight," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2324–2333, May 2016.

[10] T. L. Lai, "Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 595–608, Mar. 2000.

[11] M. Naghshvar and T. Javidi, "Sequentiality and adaptivity gains in active hypothesis testing," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 768–782, Oct. 2013.

[12] A. T. Ihler, J. W. Fisher, and A. S. Willsky, "Nonparametric hypothesis tests for statistical dependency," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2234–2249, Aug. 2004.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[14] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *J. Mach. Learn. Res.*, vol. 7, no. Jul., pp. 1283–1314, 2006.

[15] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Scholkopf, "Learning from distributions via support measure machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 10–18.

[16] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. 21st Annu. Conf. Learn. Theory*, 2008, pp. 111–122.

[17] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 473–480.

[18] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.

[19] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Proc. 30th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4240–4248.

[20] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the Grassmann manifold: Fundamental limits of subspace classifiers," in *Proc. IEEE Int. Symp. Inf. Theory*, 2014, pp. 3012–3016.

[21] M. Nokleby, A. Beirami, and R. Calderbank, "Rate-distortion bounds on Bayes risk in supervised learning," in *Proc. IEEE Int. Symp. Inf. Theory*, 2016, pp. 2099–2103.

[22] G. E. Smith, K. Woodbridge, and C. J. Baker, "Radar micro-doppler signature classification using dynamic time warping," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 3, pp. 1078–1096, Jul. 2010.

[23] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. New York, NY, USA: Springer, 2008.

[24] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, Sep. 2005.

[25] C. McDiarmid, "On the method of bounded differences," *Surv. Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[26] P. Massart, "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality," *Ann. Probability*, vol. 18, pp. 1269–1283, 1990.

Authors' photographs and biographies not available at the time of publication.