Nonparametric Detection of Anomalous Data Streams

Shaofeng Zou, *Member, IEEE*, Yingbin Liang, *Senior Member, IEEE*, H. Vincent Poor, *Fellow, IEEE*, and Xinghua Shi, *Member, IEEE*

***Abstract*—A nonparametric anomalous hypothesis testing problem is investigated, in which there are totally n observed sequences out of which s anomalous sequences are to be detected. Each typical sequence consists of m independent and identically distributed (i.i.d.) samples drawn from a distribution p, whereas each anomalous sequence consists of m i.i.d. samples drawn from a distribution q that is distinct from p. The distributions p and q are assumed to be unknown in advance. Distribution-free tests are constructed by using the maximum mean discrepancy as the metric, which is based on mean embeddings of distributions into a reproducing kernel Hilbert space. The probability of error is bounded as a function of the sample size m, the number s of anomalous sequences, and the number n of sequences. It is shown that with s known, the constructed test is exponentially consistent if m is greater than a constant factor of $\log n$, for any p and q, whereas with s unknown, m should have an order strictly greater than $\log n$. Furthermore, it is shown that no test can be consistent for arbitrary p and q if m is less than a constant factor of $\log n$. Thus, the order-level optimality of the proposed test is established. Numerical results are provided to demonstrate that the proposed tests outperform (or perform as well as) tests based on other competitive approaches under various cases.**

***Index Terms*—Anomalous hypothesis testing, consistency, distribution-free tests, maximum mean discrepancy (MMD).**

Manuscript received August 20, 2016; revised February 17, 2017 and May 26, 2017; accepted July 13, 2017. Date of publication July 31, 2017; date of current version September 5, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. D. Robert Iskander. The work of S. Zou was supported by a National Science Foundation CAREER Award under Grant CCF-10-26565, when S. Zou was a PhD student at Syracuse University. The work of Y. Liang was supported in part by a National Science Foundation CAREER Award under Grant CCF-10-26565 and in part by the National Science Foundation under Grant CCF-16-17789. The work of H. V. Poor was supported by the National Science Foundation under Grants CMMI-1435778 and ECCS-1343210. The work of X. Shi was supported by the National Science Foundation under Grants DGE-1523154 and IIS-1502172. This paper was presented in part at the 52nd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 2014. *(Corresponding author: Shaofeng Zou.)*

S. Zou is with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: szou3@illinois.edu).

Y. Liang was with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA. She is now with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: liang.889@osu.edu).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

X. Shi is with the Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: xshi3@uncc.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/TSP.2017.2733472

Fig. 1. Anomalous hypothesis testing with data sequences generated by typical distribution p and anomalous distribution q.

I. Introduction

In THIS paper, we study an anomalous hypothesis testing problem (see Fig. 1), in which there are totally n sequences out of which s anomalous sequences need to be detected. Each typical sequence consists of m independent and identically distributed (i.i.d.) samples drawn from a distribution p, whereas each anomalous sequence contains i.i.d. samples drawn from a distribution q that is distinct from p. The distributions p and q are assumed to be unknown. The goal is to build distribution-free tests to detect the s anomalous data sequences generated by q out of all data sequences.

Solutions to this problem are very useful in many applications. For example, in cognitive wireless networks, channel measurements follow different distributions p or q depending on whether the channel being measured is busy or vacant. A major issue in such networks is to identify vacant channels out of a large number of busy channels that can then be used to improve spectral efficiency. This problem was studied in [2] and [3] under the assumption that p and q are known, whereas in this paper, we study the problem with unknown p and q. Other applications include detecting anomalous events in sensor monitoring networks [4], distinguishing diseased groups with aberrant genetic markers [5], identifying differently expressed genes from gene expression profiles [6], distinguishing virus infected computers from other virus free computers [7], detecting rare objects from astronomical data that might lead to scientific discoveries [8], and distinguishing slightly modified images from other untouched images.

The parametric model of this problem has been well studied, e.g., [2], [3], in which it is assumed that p and q are known in advance. However, the nonparametric model is less explored, in which it is assumed that p and q are unknown and can be arbitrary. Recently, Li, Nitinawarat and Veeravalli proposed the nonparametric divergence-based generalized likelihood tests in [9], and characterized the error decay exponents of these tests.

However, only the case when p and q are discrete with finite alphabets was studied in [9], and their tests utilize empirical probability mass functions of p and q.

In this paper, we study the fully nonparametric model, in which p and q are arbitrary, i.e., not necessarily discrete. The major challenges to solve this problem (compared to the discrete case studied in [9]) lie in: (1) accurately estimating distributions that may be continuous with limited numbers of samples for further anomalous hypothesis testing; (2) designing low complexity tests with distributions that may be continuous; and (3) building distribution-free consistent tests and further guaranteeing exponential error decay for arbitrary distributions.

Our approach adopts the *maximum mean discrepancy (MMD)* introduced in [10] as the distance metric between two distributions. The idea is to map probability distributions into a reproducing kernel Hilbert space (RKHS) (as proposed in [11], [12]) such that the distance between the two probability distributions can be measured by the distance between their corresponding embeddings in the RKHS. MMD can be easily estimated based on samples, and hence yields low-complexity tests. In this paper, we apply MMD as a metric to construct our tests for detecting anomalous data sequences. In contrast to consistency analysis in classical theory as in [9], which assumes that the problem dimension (i.e., the number n of sequences and the number s of anomalous sequences) is fixed and the sample size m increases, our focus is on the regime in which the problem dimension (i.e., n and s) increases. This is motivated by those applications, in which anomalous sequences are required to be detected out of a large number of typical sequences. It is clear that as n becomes larger (even with fixed s), there is a greater chance that some typical sequences generated by p exhibit statistical behavior deviating from p and may be mistakenly classified as anomalous sequences. The situation with increasing s makes it even more challenging to consistently detect all anomalous sequences. It then requires that the sample size m increase correspondingly in order to guarantee accurate detection. Hence, we are interested in characterizing how the sample size m should scale with n and s in order to guarantee consistent detection.

In this paper, we adopt the following notation to express asymptotic scaling of quantities with n:

- $f(n) = O(g(n))$: there exist $k, n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq k|g(n)|$;
- $f(n) = \Omega(g(n))$: there exist $k, n_0 > 0$ s.t. for all $n > n_0$, $f(n) \geq kg(n)$;
- $f(n) = \Theta(g(n))$: there exist $k_1, k_2, n_0 > 0$ s.t. for all $n > n_0$, $k_1 g(n) \leq f(n) \leq k_2 g(n)$;
- $f(n) = o(g(n))$: for all $k > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq kg(n)$;
- $f(n) = \omega(g(n))$: for all $k > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \geq k|g(n)|$.

A. Main Contributions

We summarize our main contributions as follows.

1) We construct MMD-based distribution-free tests, which enjoy low computational complexity and are proven to be powerful for nonparametric detection.
2) We analyze performance guarantees for the proposed MMD-based tests. We bound the probability of error as a function of the sample size m, the number s of anomalous sequences, and the total number n of sequences. We then show that with s known, the constructed test is exponentially consistent if m scales at the order $\Omega(\log n)$ for any p and q, whereas with s unknown, m should scale at the order $\omega(\log n)$ (i.e., strictly larger than $\Omega(\log n)$). Thus, the lack of the information about s results in an order-level increase in sample size m needed for consistent detection.
3) We further derive a necessary condition which states that for any test to be consistent for arbitrary p and q, m needs to scale at the order $\Omega(\log n)$, and further establish the order-level optimality of the MMD-based test.
4) We provide an interesting example study, in which the distribution q is the mixture of the typical distribution p and an anomalous distribution $\tilde{q}$. In this case, anomalous sequences contain only sparse samples from the anomalous distribution. Our results for this model quantitatively characterize the impact of the sparsity level of anomalous samples on the scaling behavior of the sample size m to guarantee consistency.
5) We provide numerical results to demonstrate our theoretical assertions and compare our tests with other competitive approaches. Our numerical results demonstrate that the MMD-based test has better performance than the divergence-based generalized likelihood test proposed in [9] when the sample size m is not very large. We also demonstrate that the MMD-based test outperforms (or performs as well as) other competitive tests including the t-test, FR-Wolf test [13], FR-Smirnov test [13], Hall test [14], kernel density ratio (KDR) test [15], kernel Fisher discriminant analysis (KFDA) test [16], one-class support measure machine (OCSMM) [8] and k-means clustering [17].

B. Related Work

In this subsection, we review relevant problems and explain their differences from our model.

The parametric model of our problem with known p and q has been studied, e.g., in [2] and [3]. In fact, in [2] and [3], this problem was studied under a sequential setting which allows adaptive sampling to achieve an optimal tradeoff between the false alarm rate and the expected sample size. In this paper, we focus on the case with fixed and equal number of samples for each data stream. It is also of interest to generalize our current results to the sequential setting. As noted above, the nonparametric model with unknown p and q was studied recently in [9], where p and q are assumed to be discrete distributions. Our study addresses the general scenario in which p and q can be arbitrary (not necessarily discrete) and unknown. Furthermore, we allow the sample size to scale with the total number n of sequences (which goes to infinity), whereas [9] studies the regime in which n is fixed and only the sample size goes to infinity.

As a generalization of the classical two-sample problem, which tests whether two sets of samples are generated from the same distribution, our problem involves much richer ingredients and more technical challenges. Our problem involves the interplay of the number n of sequences, the number s of anomalous sequences, and the sample size m to guarantee test consistency, whereas the two sample problem involves only the sample com-

plexity. Furthermore, test consistency in our problem depends on the knowledge of the number of anomalous sequences, whereas the two sample problem does not have such an issue. These new issues naturally require considerably more technical effort such as analysis of the MMD estimator via samples from mixed distributions, bounding the asymptotic behavior of the difference between two MMD estimators, and development of necessary condition on sample complexity.

A type of outlier detection problem that has been widely studied in data mining, e.g., [18], [19], is that in which a number of data samples are given and outliers that are far away from other samples (typically in Euclidean distance) need to be detected. This formulation typically does not assume underlying statistic models for data samples, whereas our problem assumes that data are drawn from either p or q. Thus, our problem is to detect an outlier distribution rather than an outlier data sample.

Another related but different model has been studied in [20]–[22], which tests whether a new sample is generated from the same distribution as a given set of training samples. This problem is a binary composite hypothesis testing problem, whereas our problem involves multi-hypothesis testing, detecting anomalous sequences out of a set of sequences that contain both typical and anomalous sequences. Furthermore, this problem assumes availability of a training set of (typical) samples, whereas our problem does not assume that any sample is known to be typical in advance.

Our problem is also closely related to the group anomaly detection problem [8], [22]–[24], which is a generalization of the outlier detection problem [18], [19] with each sample being a group of data. The goal is to detect groups of data that do not conform to the behavior of the majority data samples, i.e., to detect anomalous aggregated behavior of data points out of several groups of data. This problem is related to ours in the sense that the anomaly refers to certain behavior captured by a group of data.

C. Organization of the Paper

The rest of the paper is organized as follows. In Section II, we describe the problem formulation. In Section III, we present our tests and theoretical analysis of these tests. In Section IV, we present a necessary condition to guarantee test consistency. In Section V, we provide numerical results. Finally in Section VI, we conclude the paper.

II. Problem Statement

We study an anomalous hypothesis testing problem (see Fig. 1), in which there are in total n data sequences denoted by Y_k for $1 \leq k \leq n$. Each data sequence Y_k consists of m i.i.d. samples $y_{k1}, \ldots, y_{km}$ drawn from either a typical distribution p or an anomalous distribution q, where $p \neq q$. In the sequel, we use the notation $Y_k := (y_{k1}, \ldots, y_{km})$. We assume that the distributions p and q are arbitrary and unknown in advance. Our goal is to build distribution-free tests to detect data sequences generated by q. In fact, in practice, it is quite common that typical sequences follow a single distribution p, but outlier sequences can follow multiple distributions $q_1, \ldots, q_k$. In this paper, we focus on the simple case with only a single q to present the major approach for this type of nonparametric detection problem. The tests and analysis developed here can be naturally extended to more general situations with multiple anomalous distributions.

We assume that s out of n data sequences are anomalous, i.e., are generated by the anomalous distribution q. We study both cases with s known and unknown. We are interested in the asymptotic regime, in which the number n of data sequences goes to infinity. We assume that the number s of anomalous sequences satisfies $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha \leq 1$. This includes the following three cases: (1) s is fixed, and nonzero as $n \to \infty$; (2) $s \to \infty$, but $\frac{s}{n} \to 0$ as $n \to \infty$; and (3) $\frac{s}{n}$ approaches a positive constant, which is less than or equal to 1. Some of our results are also applicable to the case with $s = 0$, i.e., the null hypothesis in which there is no anomalous sequence. We will comment on this case when the corresponding results are presented.

We next define the probability of detection error as the performance measure of tests. We let $\mathcal{I}$ denote the set that contains indices of all anomalous data sequences. Hence, the cardinality $|\mathcal{I}| = s$. We let $\hat{\mathcal{I}}^n$ denote a sequence of index sets that contain indices of all anomalous data sequences claimed by a corresponding sequence of tests.

Definition 1: A sequence of tests is consistent if

$$\lim_{n\to\infty} P_e \equiv \lim_{n\to\infty} P\{\hat{\mathcal{I}}^n \neq \mathcal{I}^n\} = 0. \tag{1}$$

We note that the above definition of consistency is with respect to the number n of sequences instead of the number m of samples. However, as n becomes large (and possibly as s becomes large), it is increasingly challenging to consistently detect all anomalous data sequences. This then requires that the number m of samples become large enough in order to more accurately detect anomalous sequences. Therefore, the limit in the above definition in fact refers to the asymptotic regime, in which m scales fast enough as n goes to infinity in order to guarantee asymptotically small probability of error.

Furthermore, for a consistent test, it is also desired that the error probability decays exponentially fast with respect to the number m of samples.

Definition 2: A sequence of tests are exponentially consistent if

$$\liminf_{m\to\infty} \left[-\frac{1}{m}\log P_e\right] \equiv \liminf_{m\to\infty} \left[-\frac{1}{m}\log P\{\hat{\mathcal{I}}^n \neq \mathcal{I}^n\}\right] > 0. \tag{2}$$

In this paper, our goal is to construct distribution-free tests to detect anomalous sequences, and characterize the scaling behavior of m with n (and possibly s) so that the developed tests are consistent (and possibly exponentially consistent).

Example with sparse anomalous samples: In this paper, we also study an interesting example, in which the distribution q is a mixture of the typical distribution p with probability $1-\epsilon$ and an anomalous distribution $\tilde{q}$ with probability ϵ, where $0 < \epsilon \leq 1$, i.e., $q = (1-\epsilon)p + \epsilon\tilde{q}$. It can be seen that if ϵ is small, the majority of samples in an anomalous sequence are drawn from the distribution p, and only sparse samples are drawn from the anomalous distribution $\tilde{q}$. The value of ϵ captures the sparsity level of anomalous samples. Here, ϵ can scale as n increases, and is hence denoted by ϵ_n. We will study how ϵ_n affects the number of samples needed for consistent detection.

III. Test and Performance Guarantee

We adopt the MMD introduced in [10] as the distance metric to construct our tests. More specifically, suppose each distribution p belonging to $\mathcal{P}$ (a set of probability distributions) is mapped to an element in the RKHS $\mathcal{H}$ as follows:

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

where $k(\cdot,\cdot)$ is the kernel function associated with $\mathcal{H}$. It was shown in [25] and [26] that the above mean embedding mapping is injective for characteristic kernels such as Gaussian and Laplace kernels. The MMD between p and q is defined to be the distance between μ_p and μ_q in the RKHS given by

$$\text{MMD}[p, q] := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{3}$$

Due to the reproducing property of the kernel, it can be shown that

$$\begin{aligned}\text{MMD}^2[p, q] = \mathbb{E}_{x,x'}[k(x, x')] - 2\mathbb{E}_{x,y}[k(x, y)] \\ + \mathbb{E}_{y,y'}[k(y, y')],\end{aligned} \tag{4}$$

where x and x' are independent and have the same distribution p, and y and y' are independent and have the same distribution q. An unbiased estimator of $\text{MMD}^2[p, q]$ based on l_1 samples of X and l_2 samples of Y is given as follows:

$$\begin{aligned}\text{MMD}_u^2[X, Y] = \frac{1}{l_1(l_1-1)} \sum_{i=1}^{l_1} \sum_{j\neq i}^{l_1} k(x_i, x_j) \\ + \frac{1}{l_2(l_2-1)} \sum_{i=1}^{l_2} \sum_{j\neq i}^{l_2} k(y_i, y_j) - \frac{2}{l_1 l_2} \sum_{i=1}^{l_1} \sum_{j=1}^{l_2} k(x_i, y_j).\end{aligned} \tag{5}$$

In this section, we design and analyze MMD-based tests for both cases with s known and unknown, respectively. We then study an example with sparse anomalous samples.

A. Known s

In this subsection, we consider the case with s known. We start with a simple case with $s = 1$, and then extend to the general case, in which $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha \leq 1$.

Consider the case with $s = 1$. For each sequence Y_k, we use $\overline{Y}_k$ to denote the $(n-1)m$ dimensional sequence that stacks all other sequences together, as given by

$$\overline{Y}_k = \{Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n\}.$$

We then compute $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ for $1 \leq k \leq n$. If Y_k is the anomalous sequence, then $\overline{Y}_k$ is fully composed of typical sequences. Hence, $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ is a good estimator of $\text{MMD}^2[p, q]$, which is a positive constant. On the other hand, if Y_k is a typical sequence, $\overline{Y}_k$ is composed of $n-2$ sequences generated by p and only one sequence generated by q. As n increases, the impact of the anomalous sequence on $\overline{Y}_k$ is negligible, and $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ is asymptotically close to zero. Based on this understanding, we construct the following test when $s = 1$. Sequence k^* is claimed to be anomalous if

$$k^* = \arg\max_{1\leq k\leq n} \text{MMD}_u^2[Y_k, \overline{Y}_k]. \tag{6}$$

The following proposition characterizes conditions under which the above test is consistent.

Proposition 1: Consider the anomalous hypothesis testing model with one anomalous sequence, i.e., $s = 1$. Suppose the test (6) applies a bounded characteristic kernel with $0 \leq k(x, y) \leq K$ for any (x, y). Then, the probability of error is upper bounded as follows:

$$P_e \leq \exp\left(\log n - \frac{m\text{MMD}^4[p, q]\left(1 - \frac{1}{n-1}\right)}{16K^2}\right). \tag{7}$$

Furthermore, the test (6) is exponentially consistent if

$$m \geq \frac{16K^2(1+\eta)}{\text{MMD}^4[p, q]} \log n, \tag{8}$$

where η is any positive constant that does not depend on any other parameters of the model.

Proof: See Appendix A. ■

Proposition 1 implies that for the scenario with one anomalous sequence, $\Omega(\log n)$ samples are sufficient to guarantee test consistency.

As we can see from Proposition 1, the choice of kernel affects the upper bound on the error probability. And the kernel should be chosen such that $\frac{\text{MMD}^2[p,q]}{K}$ is maximized. A heuristic approach is to maximize the following quantity using methods analogous to those in [27] and [28]: $\max_{i,j} \text{MMD}_u^2[Y_i, Y_j]$, which can be viewed as an empirical estimate of $\text{MMD}^2[p, q]$. In practice, we need a train-test split of the samples, i.e., splitting each sequence of samples into two groups, and then using the first group as training samples to choose the kernel and using the second group as testing samples to detect anomalous data streams.

We next consider the case with $s \geq 1$. We consider the case with $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha < \frac{1}{2}$. Although we focus on the case with $\alpha < \frac{1}{2}$, the case with $\alpha > \frac{1}{2}$ is similar, with the roles of p and q being exchanged. Our test is a natural generalization of the test (6) except that now the test chooses the sequences with the s largest values of $\text{MMD}_u^2[Y_k, \overline{Y}_k]$, which is given by

$$\begin{aligned}\hat{\mathcal{I}} = \{k : \text{MMD}_u^2[Y_k, \overline{Y}_k] \text{ is among the } s \text{ largest} \\ \text{values of } \text{MMD}_u^2[Y_i, \overline{Y}_i] \text{ for } i = 1, \ldots, n\}.\end{aligned} \tag{9}$$

The following theorem characterizes conditions under which the above test is consistent.

Theorem 1: Consider the anomalous hypothesis testing model with s anomalous sequences, where $\frac{s}{n} \to \alpha$ as $n \to \infty$ and $0 \leq \alpha < \frac{1}{2}$. Assume the value of s is known. Further assume that the test (9) applies a bounded characteristic kernel with $0 \leq k(x, y) \leq K$ for any (x, y). Then the probability of error is upper bounded as follows:

$$P_e \leq \exp\left(\log((n-s)s) - \frac{m\left(1 - \frac{2s}{n}\right)^2 \text{MMD}^4[p, q]}{16K^2\left(1 + \frac{4n-5}{(n-1)^2}\right)}\right). \tag{10}$$

Furthermore, the test (9) is exponentially consistent for any p and q if

$$m \geq \frac{16K^2(1+\eta)}{(1-2\alpha)^2 \text{MMD}^4[p,q]} \log(s(n-s)), \tag{11}$$

where η is any positive constant that does not depend on any other parameters of the model.

Proof: See Appendix B. ■

We note that $\log((n-s)s) = \Theta(\log n)$, for $1 \leq s < n$. Hence, Theorem 1 implies that even with $s > 1$ anomalous sequences, the test (9) requires only $\Omega(\log n)$ samples in each data sequence in order to guarantee test consistency. Hence, increasing s does not affect the order-level requirement on the sample size m. We further note that Theorem 1 is also applicable to the case $\alpha > \frac{1}{2}$ with the roles of p and q exchanged.

Remark 1: The computational complexity of (9) can be reduced significantly by caching the intermediate results. Consider the matrix G defined as

$$G_{k,k} = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} k(Y_{k,i}, Y_{k,j}) \text{ and } G_{k,l} = \sum_{i=1}^{m} \sum_{j=1}^{m} k(Y_{k,i}, Y_{l,j}),$$

for $1 \leq k \leq n$ and $1 \leq l \leq n$, where $G_{k,k}$ is the scaled self-similarity term in (5), and $G_{k,l}$ is the scaled cross-similarity term in (5). It can be easily verified that $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ is a linear combination of $G_{k,k}$, $\sum_{i \neq k} G_{k,i}$ and $\sum_{i \neq k} \sum_{j \neq k} G_{i,k}$. Hence, the complexity of computing $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ for every $1 \leq k \leq n$ is $O(m^2 n^2)$, which is reduced substantially.

A more computationally efficient test can be constructed using the distance metric proposed in [29] which is based on using a J-dimensional vector to represent each of the n sequences, which can be computed in $O(nmJ^3)$ time, with J typically being small. Then using the same idea as in designing (9), the total computational complexity is $O(nmJ^3)$, which is a significant improvement compared to $O(m^2 n^2)$. The techniques used in this paper can also be applied to analyze the consistency and scaling behavior of this test.

We note that Theorem 1 (which includes Proposition 1 as a special case) characterizes conditions that guarantee test consistency for a pair of fixed but unknown distributions p and q. Hence, the condition (11) depends on the underlying distributions p and q. In fact, this condition further yields the following condition that guarantees that the test will be universally consistent for arbitrary p and q.

Proposition 2 (Universal Consistency): Consider the anomalous hypothesis testing problem with s anomalous sequences. Assume that s is known. Further assume that the test (9) applies a bounded characteristic kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test (9) is universally consistent for any arbitrary pair p and q if

$$m = \omega(\log n). \tag{12}$$

Proof: This result follows from (11), $\log((n-s)s) = \Theta(\log n)$ and the fact that $\text{MMD}[p,q]$ is constant for any given p and q. ■

B. Unknown s

In this subsection, we consider the case in which s is unknown, and we focus on the scenario in which $\frac{s}{n} \to 0$ as $n \to \infty$. This includes two cases: (1) s is fixed, and (2) $s \to \infty$ and $\frac{s}{n} \to 0$ as $n \to \infty$. Without the knowledge of s, the test in (9) is not applicable, because it depends on the value of s.

In order to build a test for this case, we first observe that for each k, although $\overline{Y}_k$ contains mixed samples from p and q, it is dominated by samples from p due to the above assumption on s. Thus, for large enough m and n, $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ should be close to zero if Y_k is drawn from p, and should be far away enough from zero (in fact, close to $\text{MMD}^2[p,q]$) if Y_k is drawn from q. Based on this understanding, we construct the following test:

$$\widehat{\mathcal{I}} = \{k : \text{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\} \tag{13}$$

where $\delta_n \to 0$ and $\frac{s^2}{n^2 \delta_n} \to 0$ as $n \to \infty$. The reason for the condition $\frac{s^2}{n^2 \delta_n} \to 0$ is to guarantee that δ_n converges to 0 more slowly than $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ with Y_k drawn from p so that as n goes to infinity, δ_n asymptotically falls between $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ with Y_k drawn from p and $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ with Y_k drawn from q. We note that the scaling behavior of s as n increases needs to be known in order to choose δ_n. In practice the scale of anomalous data sequences can be estimated based on domain knowledge.

The following theorem characterizes the condition under which the test (13) is consistent.

Theorem 2: Consider the anomalous hypothesis testing problem with s anomalous sequences in which $\frac{s}{n} \to 0$ as $n \to \infty$. Assume that s is unknown in advance. Further assume that the test (13) adopts a threshold δ_n such that $\delta_n \to 0$ and $\frac{s^2}{n^2 \delta_n} \to 0$, as $n \to \infty$, and the test applies a bounded characteristic kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the probability of error is upper bounded as follows:

$$P_e \leq \exp\left(\log s - \frac{m((1-\frac{s}{n-1})^2 \text{MMD}^2[p,q] - \delta_n)^2}{8K^2\left(1+\frac{1}{(n-1)^2}\right)} \right) + \exp\left(\log(n-s) - \frac{m\left(\delta_n - \frac{s^2 \text{MMD}^2[p,q]}{(n-1)^2}\right)^2}{8K^2\left(1+\frac{1}{(n-1)^2}\right)} \right). \tag{14}$$

Furthermore, the test (13) is consistent if

$$m \geq 8(1+\eta)K^2 \max\left\{ \frac{\log(\max\{1,s\})}{\text{MMD}^4[p,q]}, \frac{\log(n-s)}{\delta_n^2} \right\}, \tag{15}$$

where η is any positive constant that does not depend on any other parameters of the model.

Proof: See Appendix C. ■

We note that Theorem 2 is also applicable to the case with $s = 0$, i.e., the null hypothesis when there is no anomalous sequence. We further note that the test (13) is not exponentially consistent. However, if $\text{MMD}[p,q]$ can be estimated from domain knowledge, we can choose the threshold $\delta_n = \frac{\text{MMD}^2[p,q]}{2}$. In this case, the test (13) is exponentially consistent if $m = \Omega(\log n)$, which can be shown similarly to Theorem 2. And it does not require knowledge of s, nor even of the scaling behavior of s.

In fact when there is no null hypothesis (i.e., $s \geq 1$), an exponentially consistent test can be built similarly as in [30]. The basic idea is to reformulate the anomalous data stream detection problem as a problem of clustering the data streams based on their generating distributions. Using this idea, we first construct two clustering centers by choosing the two data streams with the largest MMD, then assign the remaining data streams by comparing the distances to the two clustering centers. It can be shown that this test is exponentially consistent if $m = \Omega(\log n)$. This test does not require knowing s, nor even of the scaling behavior of s. However, this test depends on the assumption that $s \geq 1$ such that the two clustering centers constructed are from the typical and anomalous distributions, respectively. Hence, this test does not work if $s = 0$.

Theorem 2 implies that m should be of the order $\omega(\log n)$ to guarantee test consistency, because $\frac{s}{n} \to 0$ and $\delta_n \to 0$ as $n \to \infty$. Compared to the case with s known (for which it is sufficient for m to scale at the order $\Theta(\log n)$), the threshold on m has an order-level increase due to the lack of knowledge of s. Furthermore, the above understanding about the order-level condition on m also yields the following sufficient condition for universal test consistency.

Proposition 3 (Universal Consistency): Consider the anomalous hypothesis testing problem in which $\frac{s}{n} \to 0$ as $n \to \infty$. Assume that s is unknown in advance. Further assume that the test (13) adopts a threshold δ_n such that $\delta_n \to 0$ and $\frac{s^2}{n^2 \delta_n} \to 0$, as $n \to \infty$, and the test applies a bounded characteristic kernel with $0 \leq k(x, y) \leq K, \forall (x, y)$. Then the test (13) is universally consistent for any p and q, if

$$m = \omega(\log n). \tag{16}$$

A comparison of Proposition 3 and Proposition 2 indicates that the knowledge of s does not affect the order-level sample complexity to guarantee universal consistency.

C. Example with Sparse Anomalous Samples

We study the example with the anomalous distribution $q = (1 - \epsilon_n)p + \epsilon_n \tilde{q}$ as introduced in Section II. The following result characterizes the impact of the sparsity level ϵ_n on the scaling behavior of m to guarantee consistent detection.

Corollary 1: Consider the model with the typical distribution p and the anomalous distribution $q = (1 - \epsilon_n)p + \epsilon_n \tilde{q}$, where $0 < \epsilon_n \leq 1$. If s is known, then the test (9) is consistent if

$$m \geq \frac{16K^2(1+\eta)}{(1-2\alpha)^2 \epsilon_n^4 \mathrm{MMD}^4[p, \tilde{q}]} \log(s(n-s)), \tag{17}$$

where η is any positive constant that does not depend on any other parameters of the model.

If s is unknown, then the test (13) is consistent if

$$m \geq 16(1+\eta)K^2 \max\left\{ \frac{\log(\max\{1, s\})}{(\epsilon_n^2 \mathrm{MMD}^2[p, \tilde{q}] - \delta_n)^2}, \right.$$
$$\left. \frac{\log(n-s)}{(\delta_n - \mathbb{E}\left[\mathrm{MMD}_u^2[Y, \overline{Y}]\right])^2} \right\}, \tag{18}$$

where η is any positive constant that does not depend on any other parameters of the model, $\frac{s^2 \epsilon_n^2}{n^2 \delta_n} \to 0$ and $\frac{\delta_n}{\epsilon_n^2} \to 0$ as $n \to \infty$, Y is a sequence generated by p, and $\overline{Y}$ is a stack of $(n-1)$ sequences with s sequences generated by $\tilde{q}$ and the remaining sequences generated by p.

Proof: The proof follows from Theorems 1 and 2 by substituting as follows:

$$\begin{aligned}
&\mathrm{MMD}^2[p, q] \\
&= \mathbb{E}_{x,x'}[k(x, x')] - 2\mathbb{E}_{x,y}[k(x, y)] + \mathbb{E}_{y,y'}[k(y, y')] \\
&= \mathbb{E}_{x,x'}[k(x, x')] - 2(1 - \epsilon_n)\mathbb{E}_{x,x'}[k(x, x')] \\
&\quad - 2\epsilon_n \mathbb{E}_{x,\tilde{y}}[k(x, \tilde{y})] + (1 - \epsilon_n)^2 \mathbb{E}_{x,x'}[k(x, x')] \\
&\quad + 2\epsilon_n(1 - \epsilon_n)\mathbb{E}_{x,\tilde{y}}[k(x, \tilde{y})] + \epsilon_n^2 \mathbb{E}_{\tilde{y},\tilde{y}'}[k(\tilde{y}, \tilde{y}')] \\
&= \epsilon_n^2 \mathrm{MMD}^2[p, \tilde{q}],
\end{aligned} \tag{19}$$

where x and x' are independent with the same distribution p, y and y' are independent with the same distribution q, and $\tilde{y}$ and $\tilde{y}'$ are independent with the same distribution $\tilde{q}$. ■

Corollary 1 implies that if ϵ_n is a constant, then the scaling behavior of m needed for consistent detection does not change. However, if $\epsilon_n \to 0$ as $n \to \infty$, i.e., anomalous sequences contain more sparse anomalous samples, then m needs to scale faster with n in order to guarantee consistent detection. This is reasonable because the sample size m should have a higher order to offset the impact of the increasingly sparse anomalous samples in each anomalous sequence. Corollary 1 explicitly captures the tradeoff between the sample size m and the sparsity level ϵ_n of anomalous samples in addition to n and s.

IV. Necessary Condition and Optimality

In this section, we provide a necessary condition for any test to be consistent.

Proposition 4: Consider the anomalous hypothesis testing problem with s anomalous sequences. For any test to be consistent for arbitrary p and q, the sample size m must satisfy

$$m \geq \frac{\log n - \log 2 - 1}{D(p||q) + D(q||p)}. \tag{20}$$

Furthermore, for any test to be universally consistent, the sample size m must satisfy

$$m = \omega(\log n). \tag{21}$$

Proof: See Appendix D. ■

The sufficient and necessary conditions on sample complexity that we have derived thus far establish the following optimality of the MMD-based test.

Theorem 3 (Optimality): Consider the nonparametric anomalous hypothesis testing problem with $s \geq 1$. For s being known and unknown, the MMD-based test (9) (under the conditions in Propositions 2) and the test (13) (under the conditions in Proposition 3) are order-level optimal in sample complexity required to guarantee universal consistency for any p and q.

Proof: The proof follows by comparing Propositions 2 and 3 with Proposition 4. ■

V. Numerical Results

In this section, we provide numerical results to demonstrate our theoretical assertions, and compare our MMD-based tests with other approaches. We also apply our test to real datasets.

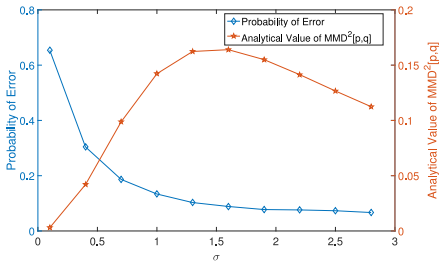


Fig. 2. Performance of the MMD-based test vs. the bandwidth parameter σ of the Gaussian kernel.

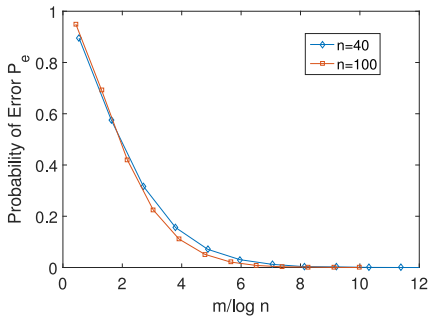


Fig. 3. The performance of the MMD-based test.

We note that although the following experiments are performed for chosen distributions p and q, our tests are nonparametric and do not exploit the information about p and q.

Previous works in kernel-based anomaly detection have shown that the Gaussian kernel is more suitable than some other kernels such as polynomial kernels [31]. Thus, we will focus on the Gaussian kernel given by $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$, where $\sigma > 0$ is the bandwidth parameter. We first study how changing σ affects the performance of our tests. We choose the distributions $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(1.2, 1)$, and set $s = 2$, $n = 10$ and $m = 20$. We plot the probability of error as a function of σ in Fig. 2. We also plot the analytically-derived value of $\text{MMD}^2[p, q]$ as a function of σ in Fig. 2. It can be observed that if σ is chosen such that the analytical value of $\text{MMD}^2[p, q]$ is maximized, the corresponding probability of error is almost minimized. This suggests that a good choice of σ can be set by maximizing the empirical estimate $\max_{1\le i,j\le n} \text{MMD}_u^2[Y_i, Y_j]$ of $\text{MMD}^2[p, q]$.

We also note that as σ increases, the probability of error approaches a constant. This is consistent with the result in [32] that for the two-sample test between distributions with different means, any bandwidth higher than a certain threshold yields equal asymptotic power. This fact will not hold for other distributions p and q. For example, if p is the Laplace distribution with mean 0 and variance 1, and $q = \mathcal{N}(0, 1)$, then the probability of error increases when σ becomes very large.

We then demonstrate our theorems on the sample complexity. We choose the distribution p to be $\mathcal{N}(0, 1)$, and choose the anomalous distribution q to be the Laplace distribution with mean one and variance one. In the experiment, we use the Gaussian kernel, and choose the bandwidth parameter σ by maximizing $\max_{1\le i,j\le n} \text{MMD}_u^2[Y_i, Y_j]$. We set $s = 1$. We run the test for cases with $n = 40$ and 100, respectively. In Fig. 3, we examine how the probability of error changes with m. For illustrational convenience, we normalize m by $\log n$, i.e., the horizontal axis represents $\frac{m}{\log n}$. It is clear that when $\frac{m}{\log n}$ is above a certain threshold, the probability of error converges to zero, which is consistent with our theoretical results. Furthermore, for these two values of n, the two curves drop to zero at almost the same threshold. This observation confirms Proposition 1, which states that the threshold on $\frac{m}{\log n}$ depends only on the bound K of the kernel and MMD of the two distributions. Both quantities are constant for the two values of n.

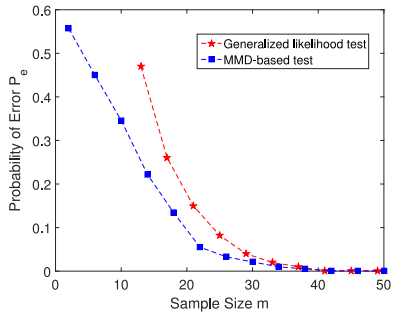


Fig. 4. Comparison of the MMD-based test with the divergence-based generalized likelihood test.

We next compare the MMD-based test with the divergence-based generalized likelihood test developed in [9]. Since the test in [9] is applicable only when the distributions p and q are discrete and have finite alphabets, we set the distributions p and q to be Bernoulli distributions with p having probability 0.3 to take "0" (and probability 0.7 to take "1"), and q having probability 0.7 to take "0" (and probability 0.3 to take "1"). We let $s = 1$ and assume that s is known. We let $n = 50$, and use the Gaussian kernel, again choosing the bandwidth parameter σ using the same method as in Fig. 3. In Fig. 4, we plot the probability of error as a function of the sample size m. It can be seen that the MMD-based test outperforms the divergence-based generalized likelihood test. We note that it has been shown in [9] that the generalized likelihood test converges at an optimal rate in the limiting case when n is infinite. Our numerical comparison, on the other hand, demonstrates that the MMD-based test performs as well as or even better than the generalized likelihood test for moderate n.

We then compare the performance of the MMD-based test with two other tests using kernel mean embedding. We choose $p = \mathcal{N}(0, 1)$, $q = \mathcal{N}(1.2, 1)$, $n = 20$, $s = 2$, and for a fair comparison, we fix the kernel to be a Gaussian kernel with $\sigma = 1$ for all three tests. We plot the probability of error as a function of the sample size m in Fig. 5 for the three algorithms: OCSMM, k-means using kernel mean embedding and our MMD-based approach.

It can be observed from Fig. 5 that MMD outperforms OCSMM and k-means. More specifically, the probability of error of OCSMM is one. This can be explained as follows. First, OCSMM does not fully exploit the knowledge of the number s of anomalous data streams. The number of anomalous data streams detected by OCSMM is not necessarily equal to the true s. Second, OCSMM detects only a subset of the anomalous sequences, and mislabels some typical sequences as anomalous ones. A possible explanation is that OCSMM is trained on the dataset in which both typical and anomalous sequences

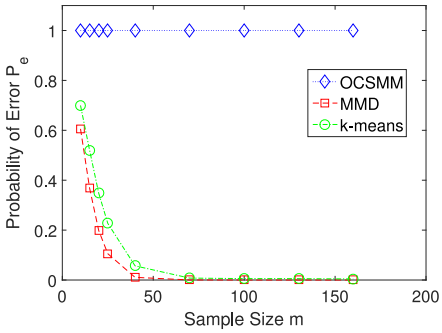


Fig. 5. Comparison of the MMD-based test with k-means and OCSMM.

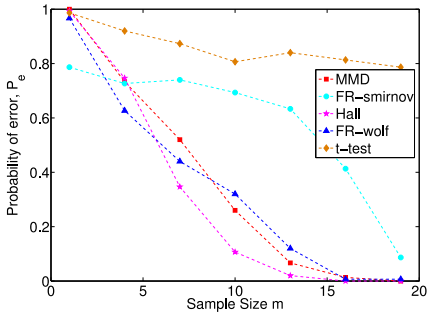


Fig. 6. Comparison of the MMD-based test with four other tests on the Syracuse-Hawaii temperature dataset.

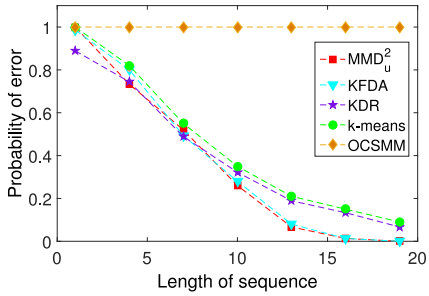


Fig. 7. Comparison of the MMD-based test with kernel-based tests on the Syracuse-Hawaii temperature dataset.

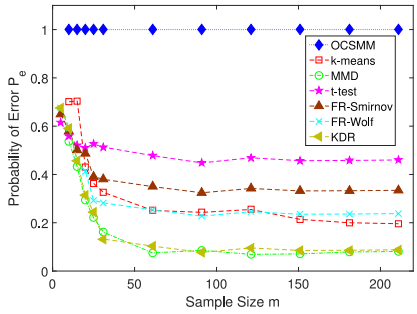


Fig. 8. Comparison of the MMD-based test with other tests on the climate-type dataset.

are mixed together, and the trained hyperplane is further used to detect the anomaly in the training data. Although it can label a large fraction of data streams correctly, it cannot detect all the anomalous sequences correctly (which is counted as an error event under our definition of the error). The k-means approach is implemented by first mapping the distributions into the RKHS, and then running the k-means algorithm using the distance in the RKHS. The k-means algorithm is usually used to solve the clustering problem and it is well-known that the k-means algorithm only performs well on balanced clusters [17].

We then compare the performance of the MMD-based test with a few other competitive tests on a Syracuse-Hawaii temperature dataset. We choose the collection of daily maximum temperatures in Syracuse (New York, USA) in July from 1993 to 2012 as the typical data sequences, and the collection of daily maximum temperatures in Makapulapai (Hawaii, USA) in May from 1993 to 2012 as anomalous sequences. Here, each data sequence contains daily maximum temperatures of a certain day across twenty years from 1993 to 2012. In our experiment, the dataset contains 32 sequences in total, including one temperature sequence from Hawaii and 31 sequences from Syracuse. The probability of error is averaged over all cases with each using one sequence from Hawaii as the anomalous sequence. Although it seems easy to detect the sequence from Hawaii out of the sequences from Syracuse, the temperatures we compare for the two places are in May for Hawaii and July for Syracuse, during which the two places have approximately the same mean temperature. Thus, it may not be easy to detect the anomalous sequence (in fact, some tests do not perform well as shown in Fig. 6).

We first compare the performance of the MMD-based test with the t-test, FR-Wolf test, FR-Smirnov test, and Hall test on the above Syracuse-Hawaii temperature dataset. For the MMD-based test, we use the Gaussian kernel with $\sigma = 1$. In Fig. 6, we plot the probability of error as a function of the sequence length m for all tests. It can be seen that the MMD-based test, Hall test, and FR-wolf test have the best performances, and all three tests are consistent with the probability of error converging to zero as m goes to infinity. Furthermore, comparing to the Hall and FR-wolf tests, the MMD-based test has the lowest computational complexity.

We further compare the performance of the MMD-based test with the kernel-based tests KFDA, KDR, OCSMM and k-means for the same dataset. For all the tests, we choose the Gaussian kernel with $\sigma = 1$ for a fair comparison. In Fig. 7, we plot the probability of error as a function of the sequence length for all tests. It can be seen that for all the tests the probability of error decreases as m increases, and the MMD-based test has the best performance among these tests.

We then compare the performance on a climate-type dataset. We obtained data taken at various weather statinos from the National Center for Atmospheric Research data archive [33]. The climate type of each station is labeled according to the Köppen-Geiger climate classification [34]. For each station, we extract the average temperate and precipitation of each month from the dataset, i.e., the dimension of each data point is two. The data at each station across months forms a sequence. We randomly choose 18 stations in southeast China and southeast North America to construct the typical sequences (191 stations in total in the chosen temperate area), and randomly choose two stations in north Africa and central Australia (13 stations

in total in the chosen tropical area) to construct the anomalous sequences. We randomly choose m months from 1987 till now, and let m vary. We plot the probability of error as a function of m in Fig. 8. It can be seen that the MMD-based approach outperforms the other approaches.

VI. Conclusion

In this paper, we have investigated a nonparametric anomalous hypothesis testing problem. We have built MMD-based distribution-free tests to detect the anomalous sequences. We have characterized the scaling behavior of the sample size m as the total number n of sequences goes to infinity in order to guarantee consistency of the developed tests. We have further provided a necessary condition for any test to be consistent, and thus established that our proposed tests are order-level optimal. Our study of this problem demonstrates a useful application of the mean embedding of distributions and MMD, and we believe that this approach can be applied to solving various other nonparametric problems.

Appendix

A. Proof of Proposition 1

We first introduce McDiarmid's inequality which is useful in bounding the probability of error in our proof.

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

$$\sup_{X \in \mathcal{X}^m, \tilde{x} \in \mathcal{X}} |f(x_1, \ldots, x_m) - f(x_1, \ldots x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)| \leq c_i. \tag{22}$$

Then for all probability distributions p and every $\epsilon > 0$,

$$P_X\left(f(X) - E_X(f(X)) > \epsilon\right) < \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right), \tag{23}$$

where X denotes $(x_1, \ldots, x_m)$, E_X denotes the expectation over the m random variables $x_i \sim p$, and P_X denotes the probability over these m variables.

In order to analyze the probability of error for the test (6), without loss of generality, we assume that the first sequence is the anomalous sequence generated by the anomalous distribution q. Hence,

$$\begin{aligned} P_e &= P(k^* \neq 1) \\ &= P\left(\exists k \neq 1 : \text{MMD}_u^2[Y_k, \overline{Y}_k] > \text{MMD}_u^2[Y_1, \overline{Y}_1]\right) \\ &\leq \sum_{k=2}^{n} P\left(\text{MMD}_u^2[Y_k, \overline{Y}_k] > \text{MMD}_u^2[Y_1, \overline{Y}_1]\right). \end{aligned} \tag{24}$$

For notational convenience, we stack $Y_1, \ldots, Y_n$ into an nm dimensional row vector $Y = \{y_i, 1 \leq i \leq nm\}$, where $Y_k = \{y_{(k-1)m+1}, \ldots, y_{km}\}$, and we define $n' = (n-1)m$. We then have,

$$\begin{aligned} \text{MMD}_u^2[Y_1, \overline{Y}_1] &= \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m,m} k(y_i, y_j) \\ &+ \frac{1}{n'(n'-1)} \sum_{\substack{i,j=m+1 \\ i \neq j}}^{nm} k(y_i, y_j) - \frac{2}{mn'} \sum_{\substack{i=1 \\ j=m+1}}^{m,nm} k(y_i, y_j). \end{aligned} \tag{25}$$

For $2 \leq k \leq n$, we have,

$$\begin{aligned} \text{MMD}_u^2[Y_k, \overline{Y}_k] &= \frac{1}{m(m-1)} \sum_{\substack{i,j=(k-1)m+1 \\ i \neq j}}^{km,km} k(y_i, y_j) \\ &+ \frac{1}{n'(n'-1)} \Bigg(\sum_{\substack{i,j=1 \\ i \neq j}}^{m,m} k(y_i, y_j) + 2 \sum_{\substack{i=1 \\ j=m+1}}^{m,(k-1)m} k(y_i, y_j) \\ &+ 2 \sum_{\substack{i=1 \\ j=km+1}}^{m,nm} k(y_i, y_j) + \sum_{\substack{i,j=m+1 \\ i \neq j}}^{(k-1)m,(k-1)m} k(y_i, y_j) \\ &+ \sum_{\substack{i,j=km+1 \\ i \neq j}}^{nm,nm} k(y_i, y_j) + 2 \sum_{\substack{i=m+1 \\ j=km+1}}^{(k-1)m,nm} k(y_i, y_j) \Bigg) \\ &- \frac{2}{mn'} \Bigg(\sum_{\substack{i=1 \\ j=(k-1)m+1}}^{m,km} k(y_i, y_j) + \sum_{\substack{i=m+1 \\ j=(k-1)m+1}}^{(k-1)m,km} k(y_i, y_j) \\ &+ \sum_{\substack{i=(k-1)m+1 \\ j=km+1}}^{km,nm} k(y_i, y_j) \Bigg). \end{aligned} \tag{26}$$

We define

$$\Delta_k = \text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_1, \overline{Y}_1].$$

It can be shown that,

$$\mathbb{E}[\text{MMD}_u^2[Y_1, \overline{Y}_1]] = \text{MMD}^2[p, q],$$

and

$$\begin{aligned} \mathbb{E}[\text{MMD}_u^2[Y_k, \overline{Y}_k]] &= \mathbb{E}_{x,x'} k(x, x') \\ &+ \frac{1}{(n-1)m((n-1)m-1)} \Big(m(m-1)\mathbb{E}_{y,y'} k(y, y') \\ &+ 2m^2(n-2)\mathbb{E}_{x,y} k(x, y) \\ &+ ((n-2)m-1)(n-2)m\mathbb{E}_{x,x'} k(x, x') \Big) \\ &- \frac{2}{(n-1)m^2} \Big(m^2 \mathbb{E}_{x,y} k(x, y) + (n-2)m^2 \mathbb{E}_{x,x'} k(x, x') \Big) \\ &= \frac{m-1}{(n-1)((n-1)m-1)} \text{MMD}^2[p, q] \end{aligned}$$

where x and x' are independent but have the same distribution p, and y and y' are independent but have the same distribution q.

We next divide the entries in $\{y_1, \ldots, y_{nm}\}$ into three groups: $Y_1 = \{y_1, \ldots, y_m\}$, $Y_k = \{y_{(k-1)m+1} \ldots, y_{km}\}$, and $\widehat{Y_k}$ that contains the remaining entries. We define Y_{-a} as Y with the a-th component y_a being removed.

For $1 \le a \le m$, y_a affects Δ_k through the following terms:

$$\frac{2}{n'(n'-1)}\Bigg(\sum_{\substack{j=1\\ j\neq a}}^{m} k(y_a, y_j) + \sum_{j=m+1}^{(k-1)m} k(y_a, y_j) + \sum_{j=km+1}^{nm} k(y_a, y_j)\Bigg) - \frac{2}{mn'}\sum_{j=(k-1)m+1}^{km} k(y_a, y_j) - \frac{2}{m(m-1)}\sum_{\substack{j=1\\ k\neq a}}^{m} k(y_a, y_j) + \frac{2}{mn'}\sum_{j=m+1}^{nm} k(y_a, y_j). \quad (27)$$

Hence, for $1 \le a \le m$, we have

$$|\Delta_k(Y_{-a}, y_a) - \Delta_k(Y_{-a}, y'_a)| \le \frac{4K}{m}\left(1 + \frac{1}{n-1}\right). \quad (28)$$

For $(k-1)m+1 \le a \le km$, y_a affects Δ_k through

$$\frac{2}{m(m-1)}\sum_{\substack{j=(k-1)m+1\\ j\neq a}}^{km} k(y_a, y_j) - \frac{2}{mn'}\Bigg(\sum_{i=1}^{m} k(y_i, y_a) + \sum_{i=m+1}^{(k-1)m} k(y_i, y_a) + \sum_{j=km+1}^{nm} k(y_a, y_j)\Bigg) - \frac{2}{n'(n'-1)}\sum_{\substack{j=m+1\\ j\neq a}}^{nm} k(y_a, y_j) + \frac{2}{mn'}\sum_{i=1}^{m} k(y_a, y_i). \quad (29)$$

Hence, for $(k-1)m+1 \le a \le km$, we have

$$|\Delta_k(Y_{-a}, y_a) - \Delta_k(Y_{-a}, y'_a)| \le \frac{4K}{m}\left(1 + \frac{1}{n-1}\right). \quad (30)$$

For $m+1 \le a \le (k-1)m$ and $km+1 \le a \le nm$, y_a affects Δ_k through

$$\frac{2}{n'(n'-1)}\Bigg(\sum_{i=1}^{m} k(y_i, y_a) + \sum_{\substack{i=m+1\\ i\neq a}}^{(k-1)m} k(y_i, y_a) + \sum_{j=km+1}^{nm} k(y_a, y_j)\Bigg) - \frac{2}{mn'}\sum_{j=(k-1)m+1}^{km} k(y_a, y_j) - \frac{2}{n'(n'-1)}\sum_{\substack{j=m+1\\ j\neq a}}^{nm} k(y_a, y_j) + \frac{2}{mn'}\sum_{i=(k-1)m+1}^{km} k(y_i, y_a). \quad (31)$$

Hence, for $m+1 \le a \le (k-1)m$ or $km+1 \le a \le nm$, we have

$$|\Delta_k(Y_{-a}, y_a) - \Delta_k(Y_{-a}, y'_a)| \le \frac{8K}{(n-1)m}. \quad (32)$$

We further derive the following probability:

$$P\Big(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \mathrm{MMD}_u^2[Y_1, \overline{Y}_1]\Big) = P\Big(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_1, \overline{Y}_1] + \mathrm{MMD}^2[p,q] - \mathbb{E}[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]] > \mathrm{MMD}^2[p,q] - \mathbb{E}[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]]\Big). \quad (33)$$

Combining (28), (30) and (32), and applying McDiarmid's inequality, we have,

$$\begin{aligned} &P\left(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \mathrm{MMD}_u^2[Y_1, \overline{Y}_1]\right) \\ &\le \exp\left(-\frac{m\mathrm{MMD}^4[p,q](1 - \frac{m-1}{(n-1)((n-1)m-1)})^2}{16K^2(1 + \frac{4n-5}{(n-1)^2})}\right) \\ &\le \exp\left(\frac{m\mathrm{MMD}^4[p,q](1 - \frac{1}{n-1})}{16K^2}\right). \end{aligned} \quad (34)$$

Hence,

$$P_e \le \exp\left(\log n - \frac{m\mathrm{MMD}^4[p,q](1 - \frac{1}{n-1})}{16K^2}\right). \quad (35)$$

We then conclude that if

$$m \ge \frac{16K^2(1+\eta)}{\mathrm{MMD}^4[p,q]}\log n, \quad (36)$$

where η is any positive constant, then $P_e \to 0$ as $n \to \infty$. It is also clear that if the above condition is satisfied, P_e converges to zero exponentially fast with respect to m. This completes the proof.

B. Proof of Theorem 1

We analyze the performance of the test (9). Without loss of generality, we assume that the first s sequences are anomalous and are generated from distribution q. Hence, the probability of error can be bounded as,

$$\begin{aligned} P_e &= P\Big(\exists k > s : \mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \min_{1\le l\le s} \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\Big) \\ &\le \sum_{k=s+1}^{n}\sum_{l=1}^{s} P\Big(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\Big). \end{aligned} \quad (37)$$

Using the fact that $\frac{s}{n} \to \alpha$, where $0 \le \alpha < \frac{1}{2}$, and using (25) and (26), we can show that for $1 \le l \le s$,

$$\begin{aligned} \mathbb{E}\left[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\right] &= \frac{(n-s)((n-s)m-1)\mathrm{MMD}^2[p,q]}{(n-1)((n-1)m-1)} \\ &\ge (1 - \frac{s}{n-1})^2\mathrm{MMD}^2[p,q], \end{aligned} \quad (38)$$

and for $s+1 \le k \le n$,

$$\mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k]\right] = \frac{s(ms-1)\text{MMD}^2[p,q]}{(n-1)((n-1)m-1)} \le \frac{s^2\text{MMD}^2[p,q]}{(n-1)^2}. \quad (39)$$

Therefore, we obtain,

$$\begin{aligned} &P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l] > 0\Big) \\ &= P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l] \\ &\quad - \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l]\right] \\ &\quad > -\mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l]\right]\Big) \\ &\le P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l] \\ &\quad - \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l]\right] \\ &\quad > \frac{(mn-1)(n-2s)\text{MMD}^2[p,q]}{(n-1)((n-1)m-1)}\Big) \\ &\le P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l] \\ &\quad - \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k] - \text{MMD}_u^2[Y_l, \overline{Y}_l]\right] \\ &\quad > (1-\frac{2s}{n})\text{MMD}^2[p,q]\Big). \end{aligned} \quad (40)$$

Applying McDiarmid's inequality, we obtain,

$$P_e \le \exp\left(\log((n-s)s) - \frac{m(1-\frac{2s}{n})^2\text{MMD}^4[p,q]}{16K^2(1+\frac{4n-5}{(n-1)^2})}\right). \quad (41)$$

Since $\frac{s}{n} \to \alpha$, as $n \to \infty$, we conclude that if

$$m \ge \frac{16K^2(1+\eta)}{(1-2\alpha)^2\text{MMD}^4[p,q]}\log(s(n-s)), \quad (42)$$

where η is any positive constant, then $P_e \to 0$, as $n \to \infty$. It is also clear that if the above condition is satisfied, P_e converges to zero exponentially fast with respect to m.

C. Proof of Theorem 2

We analyze the performance of the test (13). Without loss of generality, we assume that the first s sequences are the anomalous sequences. Hence,

$$\begin{aligned} P_e &= P\Big(\big(\exists 1 \le l \le s : \text{MMD}_u^2[Y_l, \overline{Y}_l] \le \delta_n\big)\text{or} \\ &\qquad \big(\exists s+1 \le k \le n : \text{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\big)\Big) \\ &\le \sum_{l=1}^{s} P\Big(\text{MMD}_u^2[Y_l, \overline{Y}_l] \le \delta_n\Big) \\ &\quad + \sum_{k=s+1}^{n} P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\Big). \end{aligned} \quad (43)$$

For $1 \le l \le s$, we derive,

$$\begin{aligned} &P\Big(\text{MMD}_u^2[Y_l, \overline{Y}_l] \le \delta_n\Big) \\ &= P\Big(\text{MMD}_u^2[Y_l, \overline{Y}_l] - \mathbb{E}\left[\text{MMD}_u^2[Y_l, \overline{Y}_l]\right] \\ &\qquad \le -\mathbb{E}\left[\text{MMD}_u^2[Y_l, \overline{Y}_l]\right] + \delta_n\Big) \\ &\le P\Big(\text{MMD}_u^2[Y_l, \overline{Y}_l] - \mathbb{E}\left[\text{MMD}_u^2[Y_l, \overline{Y}_l]\right] \\ &\qquad \le -(1-\frac{s}{n-1})^2\text{MMD}^2[p,q] + \delta_n\Big). \end{aligned} \quad (44)$$

For large enough n, $-(1-\frac{s}{n-1})^2\text{MMD}^2[p,q] + \delta_n < 0$.

Therefore, by applying McDiarmid's inequality, we obtain

$$\begin{aligned} &P\Big(\text{MMD}_u^2[Y_l, \overline{Y}_l] \le \delta_n\Big) \\ &\le \exp\left(-\frac{m((1-\frac{s}{n-1})^2\text{MMD}^2[p,q] - \delta_n)^2}{8K^2(1+\frac{1}{(n-1)^2})}\right), \end{aligned} \quad (45)$$

for large n.

For $s+1 \le k \le n$,

$$\begin{aligned} &P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\Big) \\ &= P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] - \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k]\right] \\ &\qquad > \delta_n - \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k]\right]\Big) \\ &\le P\Big(\text{MMD}_u^2[Y_k, \overline{Y}_k] - \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k]\right] \\ &\qquad > \delta_n - \frac{s^2\text{MMD}^2[p,q]}{(n-1)^2}\Big). \end{aligned} \quad (46)$$

Using the fact that $\frac{s^2}{n^2\delta_n} \to 0$ as $n \to \infty$, we can show that for large enough n, $\delta_n > \mathbb{E}\left[\text{MMD}_u^2[Y_k, \overline{Y}_k]\right]$. Therefore, using

McDiarmid's inequality, we have

$$P\bigg(\mathrm{MMD}_u^2[Y_k,\overline{Y}_k]>\delta_n\bigg)\le\exp\bigg(-\frac{m(\delta_n-\frac{s^2\mathrm{MMD}^2[p,q]}{(n-1)^2})^2}{8K^2(1+\frac{1}{(n-1)^2})}\bigg). \tag{47}$$

Therefore, when $s>0$,

$$\begin{aligned}P_e &\le s\exp\bigg(-\frac{m((1-\frac{s}{n-1})^2\mathrm{MMD}^2[p,q]-\delta_n)^2}{8K^2(1+\frac{1}{(n-1)^2})}\bigg)\\&\quad+(n-s)\exp\bigg(-\frac{m(\delta_n-\frac{s^2\mathrm{MMD}^2[p,q]}{(n-1)^2})^2}{8K^2(1+\frac{1}{(n-1)^2})}\bigg)\\&=\exp\bigg(\log s-\frac{m((1-\frac{s}{n-1})^2\mathrm{MMD}^2[p,q]-\delta_n)^2}{8K^2(1+\frac{1}{(n-1)^2})}\bigg)\\&\quad+\exp\bigg(\log(n-s)-\frac{m(\delta_n-\frac{s^2\mathrm{MMD}^2[p,q]}{(n-1)^2})^2}{8K^2(1+\frac{1}{(n-1)^2})}\bigg),\end{aligned} \tag{48}$$

for large enough n. Hence, we conclude that if

$$m\ge\frac{8K^2(1+\eta)}{\mathrm{MMD}^4[p,q]}\log s,\ \text{and}\ m\ge\frac{8K^2(1+\eta)}{\delta_n^2}\log(n-s),$$

where η is any positive constant, then $P_e\to 0$, as $n\to\infty$.

When $s=0$, $P_e=\sum_{k=1}^n P(\mathrm{MMD}_u^2[Y_k,\overline{Y_k}]>\delta_n)$. Then applying (47), we have that, if

$$m\ge\frac{8(1+\eta)K^2}{(\delta_n-\frac{s^2\mathrm{MMD}^2[p,q]}{(n-1)^2})^2}\log n, \tag{49}$$

where η is any positive constant, then $P_e\to 0$, as $n\to\infty$.

D. Proof of Proposition 4

Without loss of generality, we assume that $s<\frac{n}{2}$. We first construct the set of index set $\overline{\mathcal{S}}$ as follows. Let $\mathcal{I}_i^n=\{1,2,\ldots,s-1,s+i-1\}$, for $i=1,\ldots,n-s$. And let $\overline{\mathcal{S}}=\{\mathcal{I}_i^n: i=1,\ldots,n-s\}$. We note that the cardinality of each index set in $\overline{\mathcal{S}}$ is s.

Let $\overline{P}$ denote the joint distribution of $\mathcal{I}^n$ and $\{Y_1,\ldots,Y_n\}$, where $\mathcal{I}^n$ is sampled uniformly from $\overline{\mathcal{S}}$. The following Markov chain condition holds:

$$\mathcal{I}^n\to\{Y_1,\ldots,Y_n\}\to\hat{\mathcal{I}}^n \tag{50}$$

The worse-case error probability is lower bounded as follows:

$$\max_{\mathcal{I}^n:|\mathcal{I}^n|=s}P(\hat{\mathcal{I}}^n\ne\mathcal{I}^n)\ge\overline{P}(\hat{\mathcal{I}}^n\ne\mathcal{I}^n). \tag{51}$$

By Fano's inequality, and the assumption that $s<\frac{n}{2}$,

$$\begin{aligned}\overline{P}(\hat{\mathcal{I}}^n\ne\mathcal{I}^n)&\ge 1-\frac{I(\mathcal{I}^n;Y_1,\ldots,Y_n)+1}{\log|\overline{\mathcal{S}}|}\\&\ge 1-\frac{I(\mathcal{I}^n;Y_1,\ldots,Y_n)+1}{\log\frac{n}{2}},\end{aligned} \tag{52}$$

where $I(A;B)$ denotes the mutual information between A and B. Let P_i denote the distribution of $\{Y_1,\ldots,Y_n\}$ conditioned on $\mathcal{I}^n=\mathcal{I}_i^n$. Then,

$$\begin{aligned}&I(\mathcal{I}^n;Y_1,\ldots,Y_n)\\&=\sum_{i=1}^{n-s}\sum_{y_1,\ldots,y_n}P(\mathcal{I}_i^n,y_1,\ldots,y_n)\log\frac{P(\mathcal{I}_i^n,y_1,\ldots,y_n)}{P(\mathcal{I}_i^n)P(y_1,\ldots,y_n)}\\&=\sum_{i=1}^{n-s}\sum_{y_1,\ldots,y_n}\frac{1}{n-s}P_i(y_1,\ldots,y_n)\log\frac{P_i(y_1,\ldots,y_n)}{P(y_1,\ldots,y_n)},\end{aligned} \tag{53}$$

where $P(y_1,\ldots,y_n)=\frac{1}{n-s}\sum_{i=1}^{n-s}P_i(y_1,\ldots,y_n)$. Then by the Jensen's inequality and the construction of $\mathcal{I}_i^n$, it can be shown that

$$I(\mathcal{I}^n;Y_1,\ldots,Y_n)\le m(D(p||q)+D(q||p)). \tag{54}$$

Hence,

$$\overline{P}(\hat{\mathcal{I}}^n\ne\mathcal{I}^n)\ge 1-\frac{m(D(p||q)+D(q||p))+1}{\log n-\log 2}, \tag{55}$$

which implies that for any consistent test, the following condition must be satisfied:

$$m\ge\frac{(1-\eta)\log n}{D(p||q)+D(q||p)}. \tag{56}$$

Furthermore, due to the fact that $D(p||q)+D(q||p)$ is a constant for any given p and q and can be arbitrarily close to zero, for any universally consistent test, the following condition must be satisfied:

$$m=\omega(\log n). \tag{57}$$

REFERENCES

[1] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Unsupervised nonparametric anomaly detection: A kernel method," in *Proc. 52nd Allerton Conf. Commun. Control, Comput.*, Sep. 2014, pp. 836–841.

[2] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5375–5386, Aug. 2011.

[3] A. Tajer and H. V. Poor, "Quick search for rare events," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4462–4481, Jul. 2013.

[4] D. C. Harrison, W. K. G. Seah, and R. Rayudu, "Rare event detection and propagation in wireless sensor networks," *ACM Comput. Surv.*, vol. 48, no. 4, May 2016, Art. no. 58.

[5] S. Vucetic, D. Pokrajac, H. Xie, and Z. Obradovic, "Detection of underrepresented biological sequences using class-conditional distribution models," in *Proc. SIAM Int. Conf. Data Mining*, May 2003, pp. 279–283.

[6] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Data-driven approaches for detecting and identifying anomalous data streams (to appear)," in *Biomedical Signal Processing in Big Data*, T. Falk and E. Sejdic, Eds. Boca Raton, FL, USA: CRC Press, 2017.

[7] A. Honig, A. Howard, E. Eskin, and S. J. Stolfo, "System and methods for adaptive model generation for detecting intrusion in computer systems," U.S. Patent 9 497 203, Nov. 2016.

[8] K. Muandet and B. Schölkopf, "One-class support measure machines for group anomaly dection," in *Proc. Conf. Uncertain. Artif. Intell.*, Mar. 2013, pp. 449–458.

[9] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, Jul. 2014.

[10] 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, Mar. 2012.

[11] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York, NY, USA: Springer, 2004.

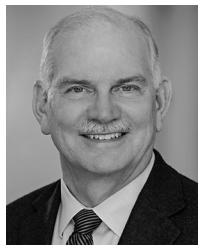
[12] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, 2010.

[13] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Stat.*, vol. 7, no. 4, pp. 697–717, 1979. [Online]. Available: http://www.jstor.org/stable/2958919
[14] P. Hall and N. Tajvidi, "Permutation tests for equality of distributions in high-dimensional settings," *Biometrika*, vol. 89, no. 2, pp. 359–374, 2002. [Online]. Available: http://www.jstor.org/stable/4140582
[15] T. Kanamori, T. Suzuki, and M. Sugiyama, "Divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 708–720, Feb. 2012.
[16] Z. Harchaoui, F. Bach, and E. Moulines, "Testing for homogeneity with kernel fisher discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 609–616.
[17] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[18] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.
[19] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
[20] A. O. Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 585–592.
[21] A. O. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 1921–1938, Sep. 1999.
[22] B. Póczos, L. Xiong, and J. Schneider, "Nonparametric divergence estimation with applications to machine learning on distributions," in *Proc. Conf. Uncertain. Artif. Intell.*, 2011, pp. 599–608.
[23] L. Xiong, B. Póczos, and J. G. Schneider, "Group anomaly detection using flexible genre models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1071–1079.
[24] B. Poczos, L. Xiong, and J. Schneider, "Nonparametric divergence estimation for learning manifolds of distributions and group anomaly detection," in *Proc. (Snowbird) Learn. Workshop*, 2011.
[25] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 489–496.
[26] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. Annu. Conf. Learn. Theory*, 2008, pp. 111–122.
[27] A. Gretton *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1205–1213.
[28] D. J. Sutherland *et al.*, "Generative models and model criticism via optimized maximum mean discrepancy," in *Proc. Int. Conf. Learn. Rep.*, Toulon, France, Apr. 2017.
[29] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton, "Fast two-sample testing with analytic representations of probability measures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1981–1989.
[30] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear complexity exponentially consistent tests for outlying sequence detection," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, Jul. 2017, pp. 983–987.
[31] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognit.*, vol. 40, no. 3, pp. 863–874, 2007.
[32] A. Ramdas, S. J. Reddi, B. Poczos, A. Singh, and L. Wasserman, "Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing," 2015, arXiv:1508.00655.
[33] N. W. S. N. U. D. o. C. Climate Prediction Center, National Centers for Environmental Prediction, "CPC global summary of day/month observations, 1979-continuing," Boulder CO, USA, 1987. [Online]. Available: http://rda.ucar.edu/datasets/ds512.0/
[34] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the Köppen-Geiger climate classification," *Hydrol. Earth Syst. Sci. Discuss.*, vol. 4, no. 2, pp. 439–473, 2007.

Shaofeng Zou (S'14–M'16) received the B.E. degree (Hons.) from Shanghai Jiao Tong University, Shanghai, China, in 2011 and the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 2016. Since July 2016, he has been a Postdoctoral Research Associate at the University of Illinois at Urbana-Champaign, Urbana, IL, USA. His research interests include information theory, machine learning, and statistical signal processing. He received the National Scholarship from the Ministry of Education of China in 2008, and the award for the outstanding graduate of Shanghai in 2011.

Yingbin Liang (S'01–M'05–SM'16) received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2005.

In 2005–2007, she was working as a Postdoctoral Research Associate at Princeton University. In 2008–2009, she was an Assistant Professor at the University of Hawaii. From 2010 to August 2017, she was an Assistant and then Associate Professor with Syracuse University, Syracuse, NY, USA. She is currently an Associate Professor with the Ohio State University, Columbus, OH, USA. Her research interests include machine learning, statistical signal processing, optimization, information theory, and wireless communication and networks. She was a Vodafone Fellow at the University of Illinois at Urbana-Champaign during 2003 to 2005, and received the Vodafone-U.S. Foundation Fellows Initiative Research Merit Award in 2005. She also received the M.E. Van Valkenburg Graduate Research Award from the ECE department, University of Illinois at Urbana-Champaign, in 2005. In 2009, she received the National Science Foundation CAREER Award, and the State of Hawaii Governor Innovation Award. In 2014, she received EURASIP Best Paper Award for the EURASIP *Journal on Wireless Communications and Networking*. She served as an Associate Editor for the Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY during 2013–2015.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in electrical engineering and computer science from Princeton University, Princeton, NJ, USA, in 1977.

From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. His research interests include the areas of information theory and signal processing, and their applications in wireless networks and related fields. Among his publications in these areas is the recent book *Information Theoretic Security and Privacy of Information Systems* (Cambridge University Press, 2017).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Royal Society. In 1990, he served as the President of the IEEE Information Theory Society, and in 2004–2007 as the Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, a D.Sc. *honoris causa* from Syracuse University awarded in 2017, and election as a Foreign Member of the National Academy of Engineering of Korea and an Honorary Member of the National Academy of Sciences of Korea, both in 2017.

Xinghua Shi (S'08–M'17) received the B.Eng. and M.Eng. degrees in computer science from Beijing Institute of Technology, Beijing, China, in 1998 and 2001, respectively, and the M.S. and Ph.D. degrees in computer science from the University of Chicago, Chicago, IL, USA, in 2003 and 2008, respectively. She obtained the Postdoctoral Trainings at Brigham and Women's Hospital and Harvard Medical School. She is an Assistant Professor in the Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, NC, USA. Her research interests include the design and development of tools and algorithms to solve large-scale computational problems in biology and network science. She is also interested in data privacy and big data analytics in biomedical research. She is an ACM life member.