Nonparametric Detection of Anomalous Data Streams

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

Abstract

A nonparametric anomalous hypothesis testing problem is investigated, in which there are totally n sequences with s anomalous sequences to be detected. Each typical sequence contains m independent and identically distributed (i.i.d.) samples drawn from a distribution p, whereas each anomalous sequence contains 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 using 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 then 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 has 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 our tests outperform (or perform as well as) the tests based on other competitive approaches under various cases.

Key words: Anomalous hypothesis testing, consistency, distribution-free tests, maximum mean discrepancy (MMD).

---

*The material in this paper was presented in part in [1] at the 52th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, Oct. 2014.

†The work of S. Zou and Y. Liang was supported by a National Science Foundation CAREER Award under Grant CCF-10-26565. The work of H. V. Poor was supported by the National Science Foundation under Grants CNS-14-56793 and ECCS-13-43210. The work of X. Shi was partly supported by National Science Foundation under Grant IIS-1502172.

‡Shaofeng Zou is with the Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana Champaign, Urbana, IL 61801 USA (email: szou3@illinois.edu). Yingbin Liang are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (email: yliang06@syr.edu). H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (email: poor@princeton.edu). Xinghua Shi is with the Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223 (email: xshi3@uncc.edu).

1 Introduction

In this paper, we study an anomalous hypothesis testing problem (see Figure 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 such a problem are very useful in many applications. For example, in cognitive wireless networks, signals follow different distributions either p or q depending on whether the channel is busy or vacant. A major issue in such a network is to identify vacant channels out of a large number of busy channels based on their corresponding signals in order to utilize vacant channels for improving 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 DNA sequences out of typical sequences, detecting virus infected computers from other virus free computers, and detecting slightly modified images from other untouched images.

The parametric model of the problem has been well studied, e.g., [2], in which it is assumed that the distributions p and q are known in advance and can be exploited for detection. However, the nonparametric model is less explored, in which it is assumed that the distributions p and q are unknown and can be arbitrary. Recently, Li, Nitinawarat and Veeravalli proposed the divergence-based generalized likelihood tests in [4], and characterized the error decay exponents of these tests. However, [4] studied only the case when the distributions p and q are discrete with finite alphabets, and their tests utilize empirical probability mass functions of p and q.

In this paper, we study the nonparametric model, in which distributions p and q can be continuous and arbitrary. The major challenges to solve this problem (compared to the discrete case studied in [4]) lie in: (1) it is difficult to accurately estimate continuous distributions with limited samples for further anomalous hypothesis testing; (2) it is difficult to design low complexity tests with continuous distributions; and (3) building distribution-free consistent tests (and further guaranteeing exponential error decay) is challenging for arbitrary distributions.

Our approach adopts the *maximum mean discrepancy (MMD)* introduced in [5] as the distance metric between two distributions. The idea is to map probability distributions into a reproducing kernel Hilbert space (RKHS) (as proposed in [6, 7]) such that the distance between the two probabilities 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 [4], 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 applications, in which anomalous sequences are required to be detected out of a large number of typical data sequences. It is clear that as n (and possibly s) becomes large, it is increasingly challenging to consistently detect all anomalous sequences. It then requires that the sample size m correspondingly increases in order to guarantee more accurate detection. Hence, we are interested in characterizing how the sample size m should scale with n and s in order to guarantee the consistency of our tests.

In this paper, we adopt the following notations 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 $n > n_0$, $|f(n)| \geq k|g(n)|$.

1.1 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 the performance guarantee for the proposed MMD-based test. 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. We further develop low complexity consistent tests by exploiting the asymptotic behavior of s and n.

(3) We further derive a necessary condition which states that no test can be consistent for arbitrary p and q if m scales at the order $O(\log n)$, thus establishing 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 distribution p and the anomalous distribution $\tilde{q}$. In such a case, the anomalous sequence contains only sparse samples from the anomalous distribution. Our results for such a model quantitatively characterize the impact of the sparsity level of anomalous samples on the scaling behavior of the sample size m, in order to guarantee consistency of the proposed tests.

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 a better performance than the divergence-based generalized likelihood test

proposed in [4] 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 t-test, FR-Wolf test [8], FR-Smirnov test [8], Hall test [9] as well as kernel density ratio (KDR) test [10] and kernel Fisher discriminant analysis (KFDA) test [11].

1.2 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]. The nonparametric model with unknown p and q were studied recently in [4], 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 [4] studies the regime in which n is fixed and only the sample size goes to infinity.

As generalization of the classical two-sample problem, which tests whether two sets of samples are generated from the same or different distributions, our problem involves much richer ingredients and more technical challenges. Our problem involves 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 complexity. 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 efforts such as analysis of the MMD estimator via samples from mixed distributions, bounding the asymptotic behavior of difference between two MMD estimators, and development of necessary conditions on sample complexity.

A popular type of outlier detection problems have been widely studied in data mining, e.g., [12, 13], 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. These studies typically do not assume underlying statistical models for data samples, whereas our problem assumes that data are drawn from either distribution 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 [14–16], which tests whether a new sample is generated from the same distribution as a given set of training samples. Such a problem is binary composite hypothesis testing, whereas our problem is multi-hypothesis testing, detecting anomalous sequences out of a set of sequences that contain both typical and anomalous sequences. Furthermore, such a problem assumes availability of a training set of (typical) samples, whereas our problem does not assume any sample known to be typical in advance.

1.3 Organization of the Paper

The rest of the paper is organized as follows. In Section 2, we describe the problem formulation. In Section 3, we present our tests and theoretical results on the performance guarantee of these tests. In Section 4, we further present the necessary conditions to guarantee test consistency. In Section 5, we provide numerical results. Finally in Section 6, we conclude our paper with remarks on future work.

2 Problem Statement

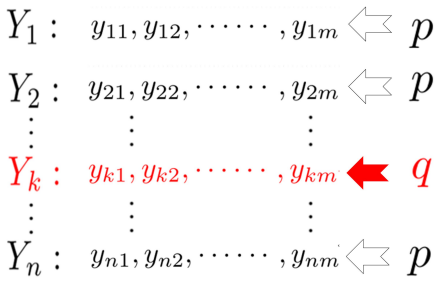

Figure 1: An anomalous hypothesis testing model with data sequences generated by typical distribution p and anomalous distribution q.

We study an anomalous hypothesis testing problem (see Figure 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 the anomalous distribution q.

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, respectively. We are interested in the asymptotical 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 to 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 such a 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 are said to be consistent if*

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

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. It then requires that the number m of samples becomes 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 desirable that the error probability decays exponentially fast with respect to the number m of samples.

Definition 2. *A sequence of tests are said to be exponentially consistent if*

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

In this paper, our goal is to construct distribution-free tests for detecting 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).

An example with sparse anomalous samples. In this paper, we also study an interesting example, in which the distribution q is a mixture of the 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 study how ϵ_n affects the number of samples needed for consistent detection.

3 Test and Performance Guarantee

We adopt the *maximum mean discrepancy (MMD)* introduced in [5] as the distance metric to construct our test. 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 has been shown in [17,18] that the above mean embedding mapping is injective for many RKHSs such as those associated with

Gaussian and Laplace kernels. The MMD between p and q is defined to be the distance between μ_p and μ_q in RKHS given by

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

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

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

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

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

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

3.1 Known s

In this subsection, we consider the case with s known. We start with a simple case with $s = 1$, and then study the more 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 $1 \leq k \leq n$. It is clear that 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 $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ should be asymptotically close to zero. Based on the above understanding, we construct the following test when $s = 1$. The 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 the condition 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 kernel with $0 \leq k(x, y) \leq K$ for any (x, y). Then, the probability of error is upper bounded as follows,*

$$P_e \leq \exp\Big(\log n - \frac{m(MMD^2[p, q] - \xi)^2}{16K^2(1 + \Theta(\frac{1}{n}))}\Big), \tag{7}$$

where ξ is a constant which can be picked arbitrarily close to zero. Furthermore, the test (6) is exponentially consistent if

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

where η is any positive constant.

Proof. See Appendix A. □

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

We next consider the case with $s \geq 1$. More specifically, 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. We first study the case with s known. Our test is a natural generalization of the test (6) except now the test picks the sequences with the largest s values of $\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]$, which is given by

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

The following theorem characterizes the condition 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 kernel with $0 \leq k(x, y) \leq K$ for any (x, y). Then the probability of error is upper bounded as follows,*

$$P_e \leq \exp\Big(\log((n - s)s) - \frac{m((1 - 2\alpha)MMD^2[p, q] - \xi)^2}{16K^2(1 + \Theta(\frac{1}{n}))}\Big), \tag{10}$$

where ξ is a constant which can be picked arbitrarily close to zero. Furthermore, the test (9) is exponentially consistent for any p and q if

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

where η is any positive constant.

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 anomalous sequence, the test (9) requires only $\Omega(\log n)$ samples in each data sequence in order to guarantee consistency of the test. Hence, the increase of 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 in which $\alpha > \frac{1}{2}$ simply with the roles of p and q exchanged.

Remark 1. *For the case with $\frac{s}{n} \to 0$, as $n \to \infty$, we can also build a test with reduced computational complexity as follows. For each Y_k, instead of using $n-1$ sequences to build $\overline{Y}_k$ as in the test (9), we take any l sequences out of the remaining $n-1$ sequences to build a sequence $\widetilde{Y}_k$, such that $\frac{l}{n} \to 0$ and $\frac{s}{l} \to 0$ as $n \to \infty$. Such an l exists for any s and n satisfying $\frac{s}{n} \to 0$ (e.g., $l = \sqrt{sn}$). It can be shown that using $\widetilde{Y}_k$ to replace $\overline{Y}_k$ in the test (9) still leads to consistent detection under the same condition given in Theorem 1. Since l is much smaller than n, computational complexity is substantially reduced.*

We note that Theorem 1 (which includes Proposition 1 as a special case) characterizes the conditions to 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, such a condition further yields the following condition that guarantees the test to be universally consistent for arbitrary p and q.

Proposition 2 (Universal Consistency)**.** *Consider the anomalous hypothesis testing problem, where $\frac{s}{n} \to \alpha$ as $n \to \infty$ and $0 \leq \alpha < \frac{1}{2}$. Assume s is known. Further assume that the test (9) applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test (9) is universally consistent for any arbitrary pair of p and q, if*

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

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

3.2 Unknown s

In this subsection, we consider the case, in which the value of s is unknown. And we focus on the scenario that $\frac{s}{n} \to 0$, 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 knowledge of s, the test in (9) is not applicable anymore, because it depends on the value of s.

In order to build a test now, 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, $\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]$ should be close to zero if Y_k is drawn from p, and should be far away enough from zero (in fact, close to $\mathrm{MMD}^2[p,q]$) if Y_k is drawn from q. Based on this understanding, we construct the following test:

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

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 $\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]$ with Y_k drawn from p so that as n goes to infinity, δ_n asymptotically falls between $\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]$ with Y_k drawn from p and $\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]$ with Y_k drawn from q. We note that the scaling behavior of s as n increases needs to be known in order to pick δ_n for the test. This is reasonable to assume because mostly 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 model with s anomalous sequences, where $\frac{s}{n} \to 0$, 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 kernel with $0 \le k(x, y) \le K$ for any (x, y). Then the probability of error is upper bounded as follows:*

$$P_e \le \exp\left(\log s - \frac{m(MMD^2[p,q] - \delta_n)^2}{16K^2(1+\Theta(\frac{1}{n}))}\right) + \exp\left(\log(n-s) - \frac{m(\delta_n - \mathbb{E}\big[MMD_u^2[Y_k, \overline{Y}_k]\big])^2}{16K^2(1+\Theta(\frac{1}{n}))}\right). \tag{14}$$

Furthermore, the test (13) *is consistent if*

$$m \ge 16(1+\eta)K^2 \max\left\{\frac{\log(\max\{1,s\})}{(MMD^2[p,q]-\delta_n)^2}\ , \frac{\log(n-s)}{(\delta_n - \mathbb{E}\big[MMD_u^2[Y,\overline{Y}]\big])^2}\right\}, \tag{15}$$

where η is any positive constant. In the above equation, $\mathbb{E}[MMD_u^2[Y, \overline{Y}]]$ is a constant, where Y is a sequence generated by p and $\overline{Y}$ is a stack of $(n-1)$ sequences with s sequences generated by q and the remaining sequences generated by p.

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. In fact, when there is no null hypothesis (i.e., $s > 1$), an exponentially consistent test can be built as follows. For each subsect $\mathcal{S}$ of $1, \ldots, n$, we compute $\mathrm{MMD}_u^2[Y_\mathcal{S}, \overline{Y}_\mathcal{S}]$, and the test finds the set of indices corresponding to the largest average value. However, for such a test to be consistent, m needs to scale linearly with n, which is not desirable.

Theorem 2 implies that m should be in 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 order-level increase due to lack of the knowledge of s. Furthermore, the above understanding on the order-level condition on m also yields the following sufficient condition for the test to be universally consistent.

Proposition 3 (Universal Consistency)**.** *Consider the anomalous hypothesis testing problem, where $\frac{s}{n} \to 0$, as $n \to \infty$. We 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 kernel with* $0 \leq k(x,y) \leq K$ *for any* (x,y)*. Then the test* (13) *is universally consistent for any arbitrary pair of* p *and* q*, if*

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

Comparison between Proposition 3 with Proposition 2 implies that the knowledge of s does not affect the order-level sample complexity to guarantee a test to be universally consistent.

3.3 Example with Sparse Anomalous Samples

We study the example with the anomalous distribution $q = (1-\epsilon_n)p + \epsilon_n\tilde{q}$ as we introduce in Section 2. The following result characterizes the impact of 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*

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

where η *is any positive constant.*

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

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

where η *is any positive constant,* $\frac{s^2\epsilon_n^2}{n^2\delta_n} \to 0$ *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:

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

where x and x' are independent but have the same distribution p, y and y' are independent but have the same distribution q, and $\tilde{y}$ and $\tilde{y}'$ are independent but have 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 cancel out the impact of the increasingly sparse anomalous samples in each anomalous sequence. Corollary 1 explicitly captures such tradeoff between the sample size m and the sparsity level ϵ_n of anomalous samples in addition to n and s.

4 Necessary Condition and Optimality

In Section 3, we characterize sufficient conditions on the sample size m under which the MMD-based test is guaranteed to be consistent for any distribution pair p and q. In this section, we characterize conditions under which no test is universally consistent for arbitrary p and q. We first study the case with $s = 1$ for which we develop our key idea of the proof. We then generalize our study to the case with $s \geq 1$.

Proposition 4. *Consider the anomalous hypothesis testing problem with one anomalous sequence. If the sample size m satisfies*

$$m = O(\log n), \tag{20}$$

then there exists no test that is universally consistent for any arbitrary distribution pair p and q.

Proof. See Appendix D. The idea of the proof is to show that for a certain distribution pair p and q, even the optimal parametric test (with known p and q) is not consistent under the condition given in the theorem. This thus implies that under the same condition, no nonparametric test is universally consistent for arbitrary p and q. □

We now generalize our result to the case with $s \geq 1$, and provide the following proposition.

Proposition 5. *Consider the anomalous hypothesis testing problem with s anomalous sequences. If the sample size m satisfies*

$$m = O\left(\frac{\log \frac{n}{s}}{s}\right), \tag{21}$$

then there exists no test that is universally consistent for arbitrary distribution pair p and q.

Proof. It can be shown that the probability of error of this problem is lower bounded by a special scenario, in which anomalous sequences can only be a group of s sequences with consecutive indices, i.e., one of the following possibilities: the $(is+1)$-th to $(i+1)s$-th sequences, for $i = 0, \ldots, \lfloor \frac{n}{s} \rfloor - 1$. Hence, there are $\lfloor \frac{n}{s} \rfloor$ candidates. Such a specific scenario can be viewed as the problem of detecting one anomalous sequence with length ms out of $\lfloor \frac{n}{s} \rfloor$ sequences. The proposition then follows from arguments similar to those used to prove Proposition 4. □

The sufficient and necessary conditions on sample complexity that we derive so far establish the following performance optimality for 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 respectively order-level optimal in sample complexity required to guarantee universal consistency for arbitrary p and q.*

Proof. The proof follows by comparing Propositions 2 and 3 with Proposition 5 and observing the fact that $m = O(\log n)$ in Proposition 5 for finite s. □

5 Numerical Results

In this section, we provide numerical results to demonstrate our theoretical assertions, and compare our MMD-based tests with a number of other tests. We also apply our test to a real data set.

We first demonstrate our theorem on sample complexity. We note that although the following experiment is performed for chosen distributions p and q, our tests are nonparametric and do not exploit the information about p and q. We choose the distribution p to be Gaussian with mean zero and variance one, i.e., $\mathcal{N}(0, 1)$, and choose the anomalous distribution q to be Laplace distribution with mean one and variance one. We use the Gaussian kernel $k(x, x') = \exp(-\frac{|x-x'|^2}{2\sigma^2})$ with $\sigma = 1$. We set $s = 1$. We run the test for cases with $n = 40$ and 100, respectively. In Figure 2, we plot 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 from the figure 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 different values of n, the two curves drop to zero almost at 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.

We next compare the MMD-based test with the divergence-based generalized likelihood test developed in [4]. Since the test in [4] is applicable only when the distributions p and q are discrete and have finite alphabets, we set the distributions p and q to be binary 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 $n = 50$. In Figure 3, 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 [4] that the generalized likelihood test has optimal convergence 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.

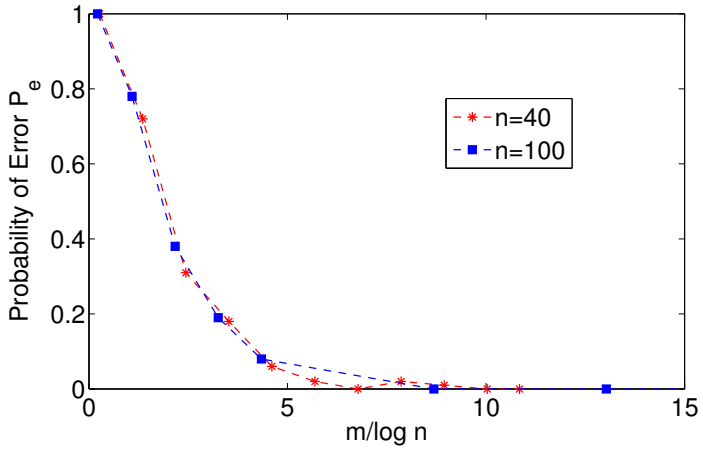

Figure 2: The performance of the MMD-based test.

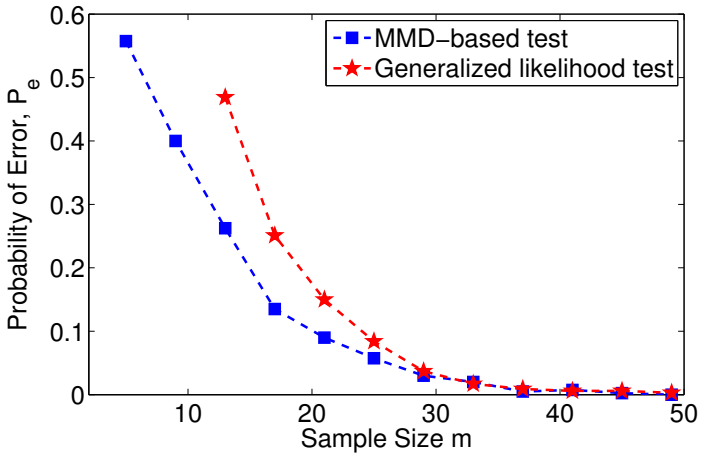

Figure 3: Comparison of the MMD-based test with divergence-based generalized likelihood test.

We finally compare the performance of the MMD-based test with a few other competitive tests on a real data set. We choose the collection of daily maximum temperature of Syracuse (New York, USA) in July from 1993 to 2012 as the typical data sequences, and the collection of daily maximum temperature of 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 data set contains 32 sequences in total, including one temperature sequence of Hawaii and 31 sequences of Syracuse. The probability of error is averaged over all cases with each using one sequence of Hawaii as the anomalous sequence. Although it seems easy to detect the sequence of Hawaii out of the sequences of 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 in temperature. In this way, it may not be easy to detect the anomalous sequence (in fact, some tests do not perform well as shown in Figure 4).

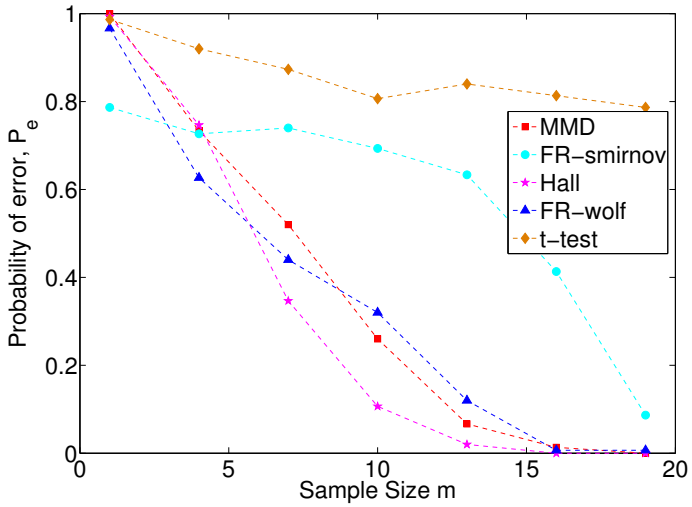

Figure 4: Comparison of the MMD-based test with four other tests on a real data set.

We first compare the performance of the MMD-based test with t-test, FR-Wolf test, FR-Smirnov test, and Hall test on the above data set. For the MMD-based test, we use the Gaussian kernel with $\sigma = 1$. In Figure 4, we plot the probability of error as a function of the length of sequence 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 of the three tests are consistent with the probability of error converging to zero as m goes to infinity. Furthermore, comparing to Hall and FR-wolf tests, the MMD-based test has the lowest computational complexity.

We further compare the performance of MMD-based test with the kernel-based tests KFDA and KDR for the same data set. For all three tests, we use Gaussian kernel with $\sigma = 1$. In Figure 5, we plot the probability of error as a function of the length of sequence for all tests. It can be seen that all tests are consistent with the probability of error converging to zero as m increases, and the MMD-based test has the best performance among the three tests.

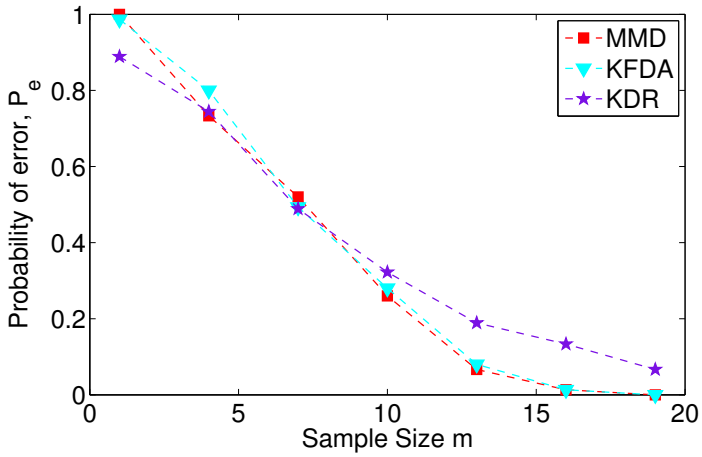

Figure 5: Comparison of the MMD-based test with two other kernel-based tests on a real data set.

6 Conclusion

In this paper, we have investigated a nonparametric anomalous hypothesis testing problem, in which typical and anomalous data sequences contain i.i.d. samples drawn from different distributions p and q, respectively. We have built MMD-based distribution-free tests to detect 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 characterized the conditions under which no test is universally consistent for arbitrary p and q, 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 such an approach can be applied to solving various other nonparametric problems.

Appendix

A Proof of Proposition 1

We first introduce the 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 measure 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,

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

For notational convenience, we stack $Y_1, \ldots, Y_n$ into a 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,

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

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

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

We define

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

It can be shown that,

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

and

$$\begin{aligned}\mathbb{E}[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]] &= \mathbb{E}_{x,x'}k(x,x') + \frac{1}{(n-1)m((n-1)m-1)}\bigg(m(m-1)\mathbb{E}_{y,y'}k(y,y') \\ &\quad + 2m^2(n-2)\mathbb{E}_{x,y}k(x,y) + ((n-2)m-1)(n-2)m\mathbb{E}_{x,x'}k(x,x')\bigg) \\ &\quad - \frac{2}{(n-1)m^2}\bigg(m^2\mathbb{E}_{x,y}k(x,y) + (n-2)m^2\mathbb{E}_{x,x'}k(x,x')\bigg) \\ &\quad \to 0, \text{ as } n \to \infty, \end{aligned} \tag{27}$$

where x and x' are independent but have the same distribution p, y and y' are independent but have the same distribution q. Hence, there exists a constant ξ that satisfies

$$\mathbb{E}[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]] < \xi < \mathrm{MMD}^2[p,q], \tag{28}$$

for large enough n. Here, ξ can be arbitrarily close to zero as $n \to \infty$.

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

$$\begin{aligned}&\frac{1}{n'(n'-1)}\bigg(2\sum_{\substack{j=1\\j\neq a}}^{m} k(y_a,y_j) + 2\sum_{j=m+1}^{(k-1)m} k(y_a,y_j) + 2\sum_{j=km+1}^{nm} k(y_a,y_j)\bigg) \\ &- \frac{2}{mn'}\sum_{j=(k-1)m+1}^{km} k(y_a,y_j) - \frac{2}{m(m-1)}\sum_{\substack{j=1\\k\neq a}}^{m} k(y_a,y_j) + \frac{2}{mn'}\sum_{j=m+1}^{nm} k(y_a,y_j). \end{aligned} \tag{29}$$

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

$$|\Delta_k(Y_{-a}, y_a) - \Delta_k(Y_{-a}, y_a')| \le \frac{4K}{m}\bigg(1 + \Theta\bigg(\frac{1}{n}\bigg)\bigg). \tag{30}$$

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

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

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

$$|\Delta_k(Y_{-a}, y_a) - \Delta_k(Y_{-a}, y_a')| \le \frac{4K}{m}\bigg(1 + \Theta\bigg(\frac{1}{n}\bigg)\bigg). \tag{32}$$

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

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

$$- \frac{2}{n'(n'-1)}\sum_{\substack{j=m+1\\ j\neq a}}^{nm} k(y_a,y_j) + \frac{2}{mn'}\sum_{i=(k-1)m+1}^{km} k(y_i,y_a). \tag{33}$$

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

$$|\Delta_k\big(Y_{-a},y_a\big) - \Delta_k\big(Y_{-a},y'_a\big)| \le \frac{1}{m}\Theta\bigg(\frac{1}{n}\bigg). \tag{34}$$

We further derive the following probability,

$$\begin{aligned}
&P\Big(\mathrm{MMD}_u^2[Y_k,\overline{Y}_k] > \mathrm{MMD}_u^2[Y_1,\overline{Y}_1]\Big)\\
&= P\Big(\mathrm{MMD}_u^2[Y_k,\overline{Y}_k] - \mathrm{MMD}_u^2[Y_1,\overline{Y}_1] + \mathrm{MMD}^2[p,q] > \mathrm{MMD}^2[p,q]\Big)\\
&\overset{(a)}{\le} P\Big(\mathrm{MMD}_u^2[Y_k,\overline{Y}_k] - \mathrm{MMD}_u^2[Y_1,\overline{Y}_1] + \mathrm{MMD}^2[p,q] - \mathbb{E}[\mathrm{MMD}_u^2[Y_k,\overline{Y}_k]] > \mathrm{MMD}^2[p,q] - \xi\Big),
\end{aligned} \tag{35}$$

where (a) follows from (28).

Combining (30), (32), (34), and applying McDiarmid's inequality, we have,

$$\begin{aligned}
&P\Big(\mathrm{MMD}_u^2[Y_k,\overline{Y}_k] > \mathrm{MMD}_u^2[Y_1,\overline{Y}_1]\Big)\\
&\quad\le \exp\bigg(-\frac{2(\mathrm{MMD}^2[p,q]-\xi)^2}{2m\frac{16K^2}{m^2}(1+\Theta(\frac{1}{n})) + \frac{1}{m}\Theta(\frac{1}{n})}\bigg)\\
&\quad= \exp\bigg(-\frac{m(\mathrm{MMD}^2[p,q]-\xi)^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg)
\end{aligned} \tag{36}$$

Hence,

$$P_e \le \exp\bigg(\log n - \frac{m(\mathrm{MMD}^2[p,q]-\xi)^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg). \tag{37}$$

Since ξ can be picked arbitrarily close to zero, we conclude that if

$$m \ge \frac{16K^2(1+\eta)}{\mathrm{MMD}^4[p,q]}\log n, \tag{38}$$

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,

$$\begin{aligned} P_e =& P\bigg(\exists k > s : \mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \min_{1\le l\le s} \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\bigg) \\ \le& \sum_{k=s+1}^{n}\sum_{l=1}^{s} P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\bigg). \end{aligned} \tag{39}$$

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

$$\mathbb{E}\big[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] \to (1-\alpha)^2\mathrm{MMD}^2[p,q], \tag{40}$$

as $n \to \infty$ for $1 \le l \le s$, and

$$\mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big] \to \alpha^2\mathrm{MMD}^2[p,q], \tag{41}$$

as $n \to \infty$ for $s+1 \le k \le n$. Hence, there exists a constant ξ such that

$$0 < \xi < (1-\alpha)^2\mathrm{MMD}^2[p,q] - \alpha^2\mathrm{MMD}^2[p,q]$$

and

$$\mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] < \alpha^2\mathrm{MMD}^2[p,q] - (1-\alpha)^2\mathrm{MMD}^2[p,q] + \xi, \tag{42}$$

for large enough n.

Therefore, we obtain,

$$\begin{aligned} & P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l] > 0\bigg) \\ =& P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l] - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] \\ & \qquad > -\mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big]\bigg) \\ \le& P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l] - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] \\ & \qquad > ((1-\alpha)^2 - \alpha^2)\mathrm{MMD}^2[p,q]) - \xi\bigg), \end{aligned} \tag{43}$$

for large enough n.

Applying McDiarmid's inequality, we obtain,

$$P_e \le \exp\left(\log((n-s)s) - \frac{m((1-2\alpha)\mathrm{MMD}^2[p,q] - \xi)^2}{16K^2(1+\Theta(\frac{1}{n}))}\right). \tag{44}$$

Since ξ can be arbitrarily small, we conclude that if,

$$m \geq \frac{16K^2(1+\eta)}{(1-2\alpha)^2 \mathrm{MMD}^4[p,q]} \log(s(n-s)), \tag{45}$$

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,

$$\begin{aligned} P_e &= P\bigg(\big(\exists 1 \leq l \leq s : \mathrm{MMD}_u^2[Y_l, \overline{Y}_l] \leq \delta_n\big) \text{or } \big(\exists s+1 \leq k \leq n : \mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\big) \bigg) \\ &\leq \sum_{l=1}^{s} P\bigg(\mathrm{MMD}_u^2[Y_l, \overline{Y}_l] \leq \delta_n \bigg) + \sum_{k=s+1}^{n} P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n \bigg). \end{aligned} \tag{46}$$

Using the fact that $\frac{s}{n} \to 0$ as $n \to \infty$, and using (25) and (26) we obtain,

$$\mathbb{E}\big[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] \to \mathrm{MMD}^2[p,q], \tag{47}$$

$$\mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big] \to 0, \tag{48}$$

as $n \to \infty$, for $1 \leq l \leq s$ and $s+1 \leq k \leq n$.

Due to (47), for any constant ϵ, $-\mathbb{E}\big[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] < -\mathrm{MMD}^2[p,q] + \epsilon$ for large enough n.

For $1 \leq l \leq s$, we drive,

$$\begin{aligned} &P\bigg(\mathrm{MMD}_u^2[Y_l, \overline{Y}_l] \leq \delta_n \bigg) \\ &= P\bigg(\mathrm{MMD}_u^2[Y_l, \overline{Y}_l] - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] \leq -\mathbb{E}\big[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l]\big] + \delta_n \bigg) \\ &\leq P\bigg(\mathrm{MMD}_u^2[Y_l, \overline{Y}_l] - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_l, \overline{Y}_l] \leq -(\mathrm{MMD}^2[p,q] - \epsilon - \delta_n) \bigg), \end{aligned} \tag{49}$$

for large enough n. Therefore, by applying McDiarmid's inequality, we obtain,

$$\begin{aligned} &P\bigg(\mathrm{MMD}_u^2[Y_l, \overline{Y}_l] \leq \delta_n \bigg) \\ &\leq \exp\bigg(-\frac{2(\mathrm{MMD}^2[p,q] - \epsilon - \delta_n)^2}{\frac{16K^2}{m}(1+\Theta(\frac{1}{n})) + \frac{16K^2}{m}(1+\Theta(\frac{1}{n}))} \bigg) \\ &= \exp\bigg(-\frac{m(\mathrm{MMD}^2[p,q] - \epsilon - \delta_n)^2}{16K^2(1+\Theta(\frac{1}{n}))} \bigg), \end{aligned} \tag{50}$$

for large enough n.

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

$$P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\bigg) \\ = P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big] > \delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big]\bigg). \tag{51}$$

Using the fact that $\frac{s^2}{n^2\delta_n} \to 0$ as $n \to \infty$, we can show that

$$\frac{\mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big]}{\delta_n} \to 0,$$

as $n \to \infty$. Hence, for large enough n, $\delta_n > \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big]$. Therefore, using McDiarmid's inequality, we have

$$P\bigg(\mathrm{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\bigg) \\ \leq \exp\bigg(-\frac{2(\delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big])^2}{\frac{16K^2}{m}(1+\Theta(\frac{1}{n})) + \frac{16K^2}{m}(1+\Theta(\frac{1}{n}))}\bigg) \\ = \exp\bigg(-\frac{m(\delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big])^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg). \tag{52}$$

Therefore,

$$P_e \leq s\exp\bigg(-\frac{m(\mathrm{MMD}^2[p,q] - \epsilon - \delta_n)^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg) \\ + (n-s)\exp\bigg(-\frac{m(\delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big])^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg) \\ = \exp\bigg(\log s - \frac{m(\mathrm{MMD}^2[p,q] - \epsilon - \delta_n)^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg) \\ + \exp\bigg(\log(n-s) - \frac{m(\delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big])^2}{16K^2(1+\Theta(\frac{1}{n}))}\bigg), \tag{53}$$

for large enough n. Hence, we conclude that if

$$m \geq \frac{16(1+\eta)K^2}{(\mathrm{MMD}^2[p,q] - \delta_n)^2}\log s, \tag{54}$$

and

$$m \geq \frac{16(1+\eta)K^2}{(\delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big])^2}\log(n-s), \tag{55}$$

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 (52), we have if

$$m \geq \frac{16(1+\eta)K^2}{(\delta_n - \mathbb{E}\big[\mathrm{MMD}_u^2[Y_k, \overline{Y}_k]\big])^2} \log n, \tag{56}$$

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

D Proof of Proposition 4

We first introduce an interesting property of Gaussian distribution, which is useful for bounding the probability of error for our problem.

Lemma 2. *[19] For the standard Gaussian distribution with mean zero and variance one, there exists positive constants c_1 and c_2 such that the cumulative distribution function (CDF) $\Phi(x)$ of the standard Gaussian distribution satisfies the following inequalities:*

$$\frac{c_1}{\log n} < \sup_{-\infty<x<\infty} |\Phi^n(a_n x + b_n) - G(x)| < \frac{c_2}{\log n} \tag{57}$$

for all positive integer n, where $G(x) = e^{e^{-x}}$ (i.e., the CDF of the Gumbel distribution), $a_n b_n = 1$. In particular, b_n can be approximated as

$$b_n = \sqrt{2\log n} - \frac{\frac{1}{2}\log(4\pi \log n)}{\sqrt{2\log n}} + O\left(\frac{1}{\log n}\right). \tag{58}$$

Our main idea of the proof is to show that under a certain distribution pair p and q, even the optimal parametric test is not consistent under the condition given in the theorem. This thus implies that under the same condition, no nonparametric test is universally consistent for arbitrary p and q. Towards this end, we consider the case, in which p and q are Gaussian with the same variance but mean shift, i.e., $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(1, 1)$. The optimal test with known p and q is the following maximum likelihood (ML) test.

$$\hat{i} = \arg\max_{1\leq i\leq n} \{P_i(Y^{nm})\}, \tag{59}$$

where $P_i(Y^{nm})$ denotes the probability of Y^{nm} if the i-th sequence is anomalous. The probability of error under the ML test is given by:

$$P_e = \frac{1}{n}\sum_{i=1}^{n} \mathcal{P}_i\Big(P_i(Y^{nm}) \leq \max_{k\neq i} P_k(Y^{nm})\Big), \tag{60}$$

where $\mathcal{P}_i$ denotes the probability evaluated when i-th sequence is anomalous. By the symmetry of the problem,

$$\mathcal{P}_i\Big(P_i(Y^{nm}) \leq \max_{k\neq i} P_k(Y^{nm})\Big) = \mathcal{P}_j\Big(P_j(Y^{nm}) \leq \max_{k\neq j} P_k(Y^{nm})\Big), \tag{61}$$

for any $1 \leq i, j \leq n$. Hence, we have

$$\begin{aligned} P_e &= \mathcal{P}_1\Big(P_1(Y^{nm}) \leq \max_{k\neq 1} P_k(Y^{nm})\Big) \\ &= \mathcal{P}_1\Big(\frac{1}{\sqrt{m}}\sum_{i=1}^{m} Y_{1i} \leq \max_{2\leq k\leq n} \frac{1}{\sqrt{m}}\sum_{i=1}^{m} Y_{ki}\Big). \end{aligned} \tag{62}$$

For convenience, we define $B_1 := \frac{1}{\sqrt{m}}\sum_{i=1}^{m} Y_{1i}$, and $B_k := \frac{1}{\sqrt{m}}\sum_{i=1}^{m} Y_{ki}$, for $2 \leq k \leq n$. Hence, $B_1 \sim \mathcal{N}(\sqrt{m}, 1)$, and $B_k \sim \mathcal{N}(0, 1)$, and they are independent from each other. With the above definitions, the probability of error can be written as

$$\begin{aligned} P_e &= \mathcal{P}\Big(B_1 \leq \max_{2\leq k\leq n} B_k\Big) \\ &= 1 - \mathcal{P}\Big(\max_{2\leq k\leq n} B_k < B_1\Big) \\ &= 1 - \mathbb{E}_B\Big\{\Phi^{n-1}(B_1)\Big\} \end{aligned} \tag{63}$$

where Φ is the CDF of B_k.

By Lemma 2, there exists a constant c independent of n, such that for all positive integer n, and for all real values x,

$$G\Big(\frac{x-b_n}{a_n}\Big) - \frac{c}{\log n} \leq \Phi^n(x) \leq G\Big(\frac{x-b_n}{a_n}\Big) + \frac{c}{\log n}, \tag{64}$$

where a_n, b_n are optimal normalizing constants, and $G(x) = e^{-e^{-x}}$ is the CDF of the Gumbel distribution.

Hence,

$$\begin{aligned} P_e &= 1 - \mathbb{E}_B\Phi^{n-1}(B_1) \\ &\geq 1 - \frac{c}{\log(n-1)} - \mathbb{E}_B\Big\{G\Big(\frac{B_1 - b_{n-1}}{a_{n-1}}\Big)\Big\} \\ &= 1 - \frac{c}{\log(n-1)} - \mathbb{E}_T\Big\{G(T)\Big\}, \end{aligned} \tag{65}$$

where $T = \frac{B_1 - b_{n-1}}{a_{n-1}}$, and $T \sim \mathcal{N}(\frac{\sqrt{m}-b_{n-1}}{a_{n-1}}, \frac{1}{a_{n-1}^2})$. The second term in (65) can be further bounded as

$$\begin{aligned} &\mathbb{E}_T\Big\{G(T)\Big\} \\ &\quad= \int_{-\infty}^{0} e^{-e^{-t}} p(t)dt + \int_{0}^{+\infty} e^{-e^{-t}} p(t)dt \\ &\quad\leq e^{-1} + P(T \geq 0) \end{aligned} \tag{66}$$

where

$$P(T \geq 0) = Q\left(\frac{0 - \frac{\sqrt{m}-b_{n-1}}{a_{n-1}}}{\frac{1}{a_{n-1}}}\right) = Q(b_{n-1} - \sqrt{m}). \tag{67}$$

In the above equations, $Q(\cdot)$ denotes the tail probability of the standard Gaussian distribution. If $m \leq 2(1-\eta)\log n$, where η is any positive constant, $b_{n-1} - \sqrt{m} \to \infty$, $Q(b_{n-1} - \sqrt{m}) \to 0$. Hence,

$$\lim_{n\to\infty} \mathbb{E}_T[G(T)] \leq e^{-1}. \tag{68}$$

Thus, with $\frac{c}{\log n} \to 0$

$$\lim_{n\to\infty} P_e \geq 1 - e^{-1} \approx 0.6321 > 0 \tag{69}$$

as $n \to \infty$. Therefore, if $m = O(\log n)$, where η is any positive constant, there exists no consistent test for any arbitrary distributions p and q.

References

[1] S. Zou, Y. Liang, H. V. Poor, and X. Shi. Unsupervised nonparametric anomaly detection: A kernel method. In *Proc. Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 836–841, 2014.

[2] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis. Quickest search over multiple sequences. *IEEE Trans. Inform. Theory*, 57(8):5375–5386, August 2011.

[3] A. Tajer and H. V. Poor. Quick search for rare events. *IEEE Trans. Inform. Theory*, 59(7):4462–4481, July 2013.

[4] Y. Li, S. Nitinawarat, and V. V. Veeravalli. Universal outlier hypothesis testing. *IEEE Trans. Inform. Theory*, 60(7):4066–4082, July 2014.

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

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

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

[8] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Ann. Statist.*, 7(4):pp. 697–717, 1979.

[9] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):pp. 359–374, 2002.

[10] T. Kanamori, T. Suzuki, and M. Sugiyama. Divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Trans. Inform. Theory*, 58(2):708–720, Feb 2012.

[11] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2008.

[12] A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Networks*, 51(12):3448–3470, August 2007.

[13] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, July 2009.

[14] A. O. Hero. Geometric entropy minimization (GEM) for anomaly detection and localization. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 585–592, 2006.

[15] A. O. Hero and O. Michel. Asymptotic theory of greedy approximations to minimal k-point random graphs. *IEEE Trans. Inform. Theory*, 45(6):1921–1938, 1999.

[16] M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 2250–2258, 2009.

[17] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2008.

[18] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proc. Annual Conference on Learning Theory (COLT)*, 2008.

[19] P. Hall. On the rate of convergence of normal extremes. *Journal of Applied Probability*, 16(2):433–439, 1979.