NONPARAMETRIC DETECTION OF AN ANOMALOUS DISK OVER A TWO-DIMENSIONAL LATTICE NETWORK

Shaofeng Zou[⋆] *Yingbin Liang* [⋆] *H. Vincent Poor* [†]

[⋆] Syracuse University [†] Princeton University
Email: szou02@syr.edu, yliang06@syr.edu, poor@princeton.edu

ABSTRACT

Nonparametric detection of existence of an anomalous disk over a lattice network is investigated. If an anomalous disk exists, then all nodes belonging to the disk observe samples generated by a distribution q, whereas all other nodes observe samples generated by a distribution p that is distinct from q. If there does not exist an anomalous disk, then all nodes receive samples generated by p. The distributions p and q are arbitrary and unknown. The goal is to design statistically consistent test as the network size becomes asymptotically large. A kernel-based test is proposed based on maximum mean discrepancy (MMD) which measures the distance between mean embeddings of distributions into a reproducing kernel Hilbert space (RKHS). A sufficient condition on the minimum size of candidate anomalous disks is characterized in order to guarantee the consistency of the proposed test. A necessary condition that any universally consistent test must satisfy is further derived. Comparison of sufficient and necessary conditions yields that the proposed test is order-level optimal.

Index Terms— Consistency, maximum mean discrepancy, nonparametric detection, reproducing kernel Hilbert space.

1. INTRODUCTION

We study the problem of detecting existence of an anomalous disk over a lattice network, in which each node observes a random sample. If an anomalous disk exists, then all nodes belonging to the disk take samples generated by an anomalous distribution q. All other nodes in the network take samples generated by a typical distribution p. It is assumed that p and q are distinct. If there does not exist an anomalous disk, then all nodes receive samples generated by p. We assume that the distributions p and q are *unknown a priori*, and hence the problem is nonparametric. This is a composite hypothesis testing problem, in which the null hypothesis corresponds to the case with no anomalous disk and the alternative hypothesis corresponds to the case with existence of an anomalous disk. The alternative hypothesis is composite because the anomalous disk can be one of a number of candidate disks in the network.

Detecting existence of an anomalous geometric structure in large networks has been extensively studied in the literature. Most previous studies [1–10] have adopted *parametric* or *semiparametric* models, which assume that samples are generated by known distributions such as Gaussian or Bernoulli distributions, or the two distributions are known to be different by a mean shift. However, such parametric models may not always hold in real applications because distributions may not be known in advance, or even structures of distributions may not be learned. More recently, in [11], a nonparametric problem of detecting existence of an interval over a line network was studied. Although it is assumed that distributions are unknown in [11], a reference sequence of samples generated from the typical distribution is assumed to be available. The problem is easier with such an identified reference sequence.

In contrast to previous studies, we study the *nonparametric* problem, in which not only distributions are *unknown a priori*, but no reference samples are available either. We study the problem of detecting the existence of a disk over a two-dimensional network, and the network structure is more complicated than the line network in [11]. Our study provides further understanding of the impact of geometric structure on detection performance.

In our problem, the distributions are unknown, and only samples generated by the distributions are available. It is desirable to find a way to measure the distance between distributions based on samples. Towards this end, mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [12, 13] is useful. The idea is to map probability distributions into an RKHS such that the distance between two probability distributions can be measured by the distance between the corresponding mean embeddings in the RKHS. The main advantage of such an approach is that the mean embedding of a distribution can be easily estimated based on samples. This approach has been applied to solving the two sample problem in [14], in which the quantity of *maximum mean*

The work of S. Zou and Y. Liang was supported by an NSF CAREER Award under Grant CCF-10-26565. The work of H. V. Poor was supported by the ARO under MURI Grant W911NF-11-1-0036 and by the NSF under Grants CNS-14-56793 and EECS-13- 43210.

discrepancy (MMD) was used as a metric of distance between mean embeddings of two distributions. In [11], MMD was used to develop tests for detection of existence of anomalous intervals over a line network. In this paper, we further generalize the MMD-based approach to studying a more difficult anomaly detection problem without reference samples and with a more complicated network structure. Our study necessarily involves new analysis of the performance due to such generality beyond that in [11].

We assume that the network is an n-by-n lattice. We are interested in the asymptotic scenario in which the network size goes to infinity, i.e., $n \to \infty$, and the number of candidate anomalous disks scales with n. Thus, the number of sub-hypotheses under the alternative hypothesis also increases, which causes the composite hypothesis testing problem to be more difficult. As n becomes large, it is necessary that the numbers of samples within and outside of each candidate anomalous disk scale with n fast enough in order to provide more accurate information about both distributions p and q and guarantee asymptotically small probability of error. Thus, the problem amounts to characterizing how the minimum and maximum sizes of all candidate anomalous disks should scale with n in order to accurately detect the existence of an anomalous structure, where the size of a disk is defined to be the number of nodes contained in the disk.

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

- $f(n) = \Omega(g(n))$: there exist $c, n_0 > 0$ s.t. for all $n > n_0$, $f(n) \geq cg(n)$;
- $f(n) = \Theta(g(n))$: there exist $c_1, c_2, n_0 > 0$ s.t. for all $n > n_0$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$;
- $f(n) = \omega(g(n))$: for all $c > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \geq c|g(n)|$.

Our main contribution lies in comprehensive analysis of the performance guarantee for the MMD-based tests that we propose to solve the problem. We show that as n goes to infinity (i.e., the network size becomes large), if the minimum size $D_{\min}$ of candidate anomalous disks scales as $\Omega(\log n)$, then the proposed test is consistent. There is no condition on the maximum size $D_{\max}$ of candidate anomalous disks because even the largest disk cannot fully cover the entire lattice and it can be shown that samples outside the largest disk are sufficient to provide information about the distribution p. We further derive a necessary condition on $D_{\min}$ that any test must satisfy in order to be universally consistent for arbitrary p and q. Comparison of sufficient and necessary conditions yields that the MMD-based test is order-level optimal.

2. PROBLEM STATEMENT AND PRELIMINARIES

2.1. Problem statement

We consider a two-dimensional lattice network (see Figure 1) consisting of n^2 nodes placed at the corner points of a lattice.

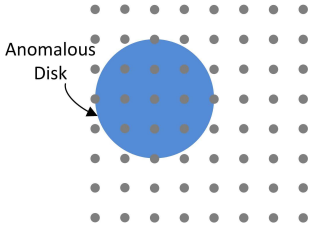

Fig. 1. A two dimensional network with an anomalous disk.

We use D to denote a subset of nodes within a *disk*. Here, the size of a disk D refers to the cardinality of D, and is denoted by $|D|$. Consider the following set of candidate anomalous disks, which consists of all disks centered at a certain node with integer radius:

$$\mathcal{D}_n^{(a)} = \{D : D_{\min} \leq |D| \leq D_{\max}\}, \tag{1}$$

where $|D|$ denotes the number of nodes within the disk D, $D_{\min}$ denotes the minimum number of nodes among the candidate anomalous disks and $D_{\max}$ denotes the maximum number of nodes among the candidate anomalous disks.

We assume that each node i observes a random sample Y_i, for $i = 1, \ldots, n^2$. Under the *null hypothesis* H_0, Y_i for $i = 1, \ldots, n^2$ are independent and identically distributed (i.i.d.) following a distribution p. Under the *alternative hypothesis* H_1, there exists a disk $D \in \mathcal{D}_n^{(a)}$ over which Y_i (with $i \in D$) are i.i.d. following a distribution $q \neq p$, and otherwise, Y_i are i.i.d. following the distribution p. Thus, the alternative hypothesis is composite due to the fact that $\mathcal{D}_n^{(a)}$ contains multiple candidate anomalous disks, and these disks differentiate from each other by their sizes and locations in the network. We further assume that under both hypotheses, each node observes only one sample.

We assume that the distributions p and q are *arbitrary* and *unknown a priori*. For this problem, we are interested in the asymptotic scenario, in which the number of nodes goes to infinity, i.e., $n \to \infty$. The performance of a test for such a system is captured by two types of errors. The *type I error* refers to the event that samples are generated from the null hypothesis, but the detector determines that an anomalous disk exists. We denote the probability of such an event as $P(H_1|H_0)$. The *type II error* refers to the case that an anomalous disk exists but the detector claims that there is no anomalous disk. We denote the probability of such an event as $P(H_0|H_1)$. We define the following risk to measure the performance of a test:

$$R_m^{(n)} = P(H_1|H_0) + \max_{D \in \mathcal{D}_n^{(a)}} P(H_0|H_{1,D}). \tag{2}$$

Definition 1. *A test is said to be consistent if the risk* $R_m^{(n)} \to 0$, *as* $n \to \infty$.

2.2. Preliminaries of MMD

We provide a brief introduction of the idea of mean embedding of distributions into an RKHS [12, 13] and the metric of MMD. Suppose $\mathcal{P}$ includes a class of probability distributions, and suppose $\mathcal{H}$ is the RKHS with an associated kernel $k(\cdot,\cdot)$. We define a mapping from $\mathcal{P}$ to $\mathcal{H}$ such that each distribution $p \in \mathcal{P}$ is mapped into an element in $\mathcal{H}$ as follows:

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

Here, $\mu_p(\cdot)$ is referred to as the *mean embedding* of the distribution p into the Hilbert space $\mathcal{H}$. Due to the reproducing property of $\mathcal{H}$, it is clear that $\mathbb{E}_p[f] = \langle \mu_p, f\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

It is desirable that the embedding is *injective* such that each $p \in \mathcal{P}$ is mapped to a unique element $\mu_p \in \mathcal{H}$. It has been shown in [13] and [15–17] that for many RKHSs such as those associated with Gaussian and Laplacian kernels, the mean embedding is injective. In order to distinguish between two distributions p and q, [14] introduced the following quantity based on the mean embeddings μ_p and μ_q of p and q in the RKHS:

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

It is also shown that

$$\text{MMD}[p, q] = \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)].$$

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

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

where x and x' are independent but have the same distribution p, and y and y' are independent but have the same distribution q. An unbiased estimate of $\text{MMD}^2[p, q]$ based on n samples of x and m samples of y is given by

$$\begin{aligned}\text{MMD}^2_u[X, Y] =& \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\ne i}^{n} k(x_i, x_j) \\ &+ \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j\ne i}^{m} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j).\end{aligned} \tag{4}$$

And it can be show that $\mathbb{E}[\text{MMD}^2_u[X, Y]] = \text{MMD}^2[p, q]$.

3. MAIN RESULTS

3.1. Test and performance

We construct a nonparametric test using the unbiased estimator in (4) and the scan statistics. For each disk D, let Y_D denote the samples in the disk D, and $Y_{\overline{D}}$ denote the samples outside the disk D. We compute $\text{MMD}^2_{u,D}(Y_D, Y_{\overline{D}})$ for all disks $D \in \mathcal{D}_n^{(a)}$. Under the null hypothesis H_0, all samples are generated from the distribution p. Hence, for each $D \in \mathcal{D}_n^{(a)}$, $\text{MMD}^2_{u,D}(Y_D, Y_{\overline{D}})$ yields an estimate of $\text{MMD}^2[p, p]$, which is zero. Under the alternative hypothesis H_1, there exists an anomalous disk D^* in which the samples are generated from distribution q. Hence, $\text{MMD}^2_{u,D^*}(Y_{D^*}, Y_{\overline{D^*}})$ yields an estimate of $\text{MMD}^2[p, q]$, which is bounded away from zero due to the fact that $p \ne q$. Based on the above understanding, we build the following test:

$$\max_{D:D\in\mathcal{D}_n^{(a)}} \text{MMD}^2_{u,D}(Y_D, Y_{\overline{D}}) \begin{cases} \ge t, & \text{determine } H_1 \\ < t, & \text{determine } H_0 \end{cases} \tag{5}$$

where t is a threshold parameter. It is anticipated that with a sufficiently accurate estimate of MMD and an appropriate choice of the threshold t, the test (5) should provide desired performance. We can further reduce the complexity of the test (5) by the fast multiscale methods in [1]. The following theorem characterizes the performance of the proposed test.

Theorem 1. *Suppose the test* (5) *is applied to the nonparametric problem described in Section 2.1. Further assume that the kernel in the test satisfies* $0 \le k(x, y) \le K$ *for all* (x, y). *Then, the type I error is upper bounded as follows:*

$$\begin{aligned}P(H_1|H_0) \le \exp\Bigg(& 3\log n \\ &- \frac{t^2 \min\{D_{\min}(n^2 - D_{\min}), D_{\max}(n^2 - D_{\max})\}}{8n^2K^2}\Bigg),\end{aligned} \tag{6}$$

and the type II error is upper bounded bounded as follows:

$$\begin{aligned}P(H_0|H_{1,D}) \le& \exp\left(-\frac{(\textit{MMD}^2[p, q] - t)^2|D|(n^2 - |D|)}{8n^2K^2}\right), \\ &\textit{for any } D \in \mathcal{D}_n^{(a)}\end{aligned} \tag{7}$$

where t *is the threshold of the test that satisfies* $t < \textit{MMD}^2[p, q]$.

Furthermore, the test (5) *is consistent if*

$$D_{\min} \ge \frac{24K^2(1 + \eta)}{t^2} \log n, \tag{8}$$

where η *is any positive constant.*

The proof of the above theorem is omitted due to space limitations. The detailed proof can be found in [18].

The above theorem implies that to guarantee consistency of the proposed test, the minimum size $D_{\min}$ should scale on the order of $\Omega(\log n)$. In fact, it should also be required that $n^2 - D_{\max}$ scale on the order of $\Omega(\log n)$ for large enough n. However, the largest disk within a two-dimensional lattice network has radius $\frac{n}{2}$ and area $\frac{\pi n^2}{4} \approx 0.79n^2$, which contains at most cn^2 nodes with $c < 1$ for large n. This implies that the bound on $D_{\max}$ is satisfied automatically when n is large. We

further note that the above theorem implies that the number of candidate anomalous disks in the set $\mathcal{D}_n^{(a)}$ is on the order of $\Theta(n^3)$, which is the same as the number of all disks. Hence, at the order level, not many disks are excluded from being anomalous.

Theorem 1 requires that the threshold t in the test (5) be less than $\text{MMD}^2[p,q]$. In fact, information of $\text{MMD}^2[p,q]$ may or may not be available depending on specific applications. In some cases, samples from anomalous events are also collected, and hence $\text{MMD}^2[p,q]$ can be estimated reasonably well by (4). In such cases, the threshold t can be set as a constant smaller than $\text{MMD}^2[p,q]$. On the other hand, if samples from q are not available, then the threshold t needs to scale to zero as n gets large in order to be asymptotically smaller than $\text{MMD}^2[p,q]$. We summarize these two cases in the following corollaries.

Corollary 1. *If the value $MMD^2[p,q]$ is known a priori, we set the threshold $t = (1-\delta)MMD^2[p,q]$ for any $0 \leq \delta \leq 1$. The test* (5) *is consistent, if $D_{\min}$ satisfies the following condition:*

$$D_{\min} \geq \frac{24K^2(1+\eta')}{MMD^4[p,q]} \log n, \tag{9}$$

where η' is any positive constant.

Corollary 1 follows directly from Theorem 1 by setting $\eta' = \frac{1+\eta}{(1-\delta)^2} - 1$.

Corollary 2. *If the value $MMD^2[p,q]$ is unknown, we set the threshold t to scale with n, such that $\lim_{n\to\infty} t_n = 0$. The test* (5) *is consistent, if $D_{\min}$ satisfies the following condition:*

$$D_{\min} \geq \frac{24K^2(1+\eta)}{t_n^2} \log n, \tag{10}$$

where η is any positive constant.

Corollary 2 follows directly from Theorem 1 by noting that $t_n < \text{MMD}^2[p,q]$ for large n.

We note that the above two corollaries demonstrate that the prior knowledge about $\text{MMD}^2[p,q]$ is very important to determine the capability for identifying anomalous events. If $\text{MMD}^2[p,q]$ is known, then an anomalous object at size $\Omega(\log n)$ can be resolved over the network. However, if such knowledge is unknown, only bigger anomalous objects at size $\omega(\log n)$ can be resolved.

Moreover, we would like to study the conditions under which our test (5) is universally consistent, i.e., consistent for any arbitrary p and q. Such conditions should not depend on the underlying p and q.

Corollary 3. *Suppose the test in* (5) *is applied to the nonparametric problem described in Section 2.1. The test* (5) *is universally consistent for any arbitrary p and q if*

$$D_{\min} = \omega(\log n). \tag{11}$$

Proof. If $D_{\min}$ satisfies the condition (11), it also satisfies the conditions (8), (9) and (10) with t_n properly chosen for any p and q when n is large enough. □

3.2. Necessary condition and optimality

In Section 3.1, we characterize a sufficient condition on $D_{\min}$ to guarantee that the proposed nonparametric test will be consistent. In the following theorem, we give a necessary conditions on $D_{\min}$ that any test must satisfy in order to be universally consistent for arbitrary p and q.

Theorem 2. *For the nonparametric detection problem described in Section 2.1, any test must satisfy the following condition in order to be universally consistent for arbitrary p and q:*

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

Outline of the Proof. The idea of the proof is to first lower bound the risk by the Bayes risk of a simpler problem. Then for such a problem, we show that there exist p and q (in fact Gaussian p and q) such that even the optimal parametric test is not consistent if the condition (12) is not satisfied. This thus implies that if (12) is not satisfied, no nonparametric test is universally consistent for arbitrary p and q. The detailed proof can be found in [18]. □

It can be seen that the necessary condition on $D_{\min}$ in (12) matches the sufficient condition in (11) at the order level, which implies that the proposed test is order-level optimal as stated in the following theorem.

Theorem 3 (Optimality)**.** *Consider the problem of nonparametric detection described in Section 2.1. The MMD-based test* (5) *is order-level optimal in the terms of the minimum size of all candidate disks required to guarantee universal test consistency for arbitrary p and q.*

4. CONCLUSION

We have studied the nonparametric detection of the existence of an anomalous disk over a two dimensional lattice network, in which both the typical and the anomalous distributions can be arbitrary and unknown. We have developed a nonparametric test using the MMD to measure the distance between the mean embeddings of distributions into an RKHS. We have analyzed the performance of our test, and provided a sufficient condition on the minimum size of candidate anomalous disks to guarantee the consistency of our test. We have further derived a necessary condition on the minimum size of candidate anomalous disks, which matches the sufficient condition at the order level. This implies that our test is order-level optimal. We believe that such an approach can be applied to study other networks such as detecting the existence of an anomalous rectangle in *r*-dimensional lattice networks, etc.

5. REFERENCES

[1] E. Arias-Castro, D. L. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Trans. Inform. Theory*, vol. 51, no. 7, pp. 2402–2425, July 2005.

[2] G. Walther, "Optimal and fast detection of spatial clusters with scan statistics," *Ann. Statist.*, vol. 38, no. 2, pp. 1010–1033, 2010.

[3] P. M. Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman, "False discovery control for random fields," *J. Amer. Stat. Assoc.*, vol. 99, pp. 1002–1014, 2004.

[4] E. Arias-Castro, E. J. Candes, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *Ann. Statist.*, vol. 36, no. 4, pp. 1726–1757, 2008.

[5] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, "On combinatorial testing problems," *Ann. Statist.*, vol. 38, no. 5, pp. 3063–3092, 2010.

[6] E. Arias-Castro, E. J. Candes, and A. Durand, "Detection of an anomalous cluster in a network," *Ann. Statist.*, vol. 39, no. 1, pp. 278–304, 2011.

[7] J. Sharpnack, A. Rinaldo, and A. Singh, "Changepoint detection over graphs with the spectral scan statistic," in *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, May 2013.

[8] J. Sharpnack, A. Rinaldo, and A. Singh, "Detecting activations over graphs using spanning tree wavelet bases," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, May 2013.

[9] J. Qian, V. Saligrama, and Y. Chen, "Connected subgraph detection," in *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, 2014, pp. 796–804.

[10] J. Qian and V. Saligrama, "Efficient minimax signal detection on graphs," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2708–2716.

[11] S. Zou, Y. Liang, and H. V. Poor, "A kernel-based nonparametric test for anomaly detection over line networks," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

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

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

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

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

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

[17] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2009.

[18] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of geometric structures over networks," *available at http://szou02.mysite.syr.edu/anomalydetection.pdf*, 2015.