Universal Outlier Hypothesis Testing

Yun Li, Sirin Nitinawarat and Venugopal V. Veeravalli
Department of Electrical and Computer Engineering
and
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801-2307, USA
Emails: yunli2@illinois.edu, nitinawa@illinois.edu, vvv@illinois.edu

***Abstract*—The following outlier hypothesis testing problem is studied in a universal setting. Vector observations are collected each with $M \geq 3$ coordinates. When a given coordinate is the outlier, the observations in that coordinate are assumed to be distributed according to the "outlier" distribution, distinct from the common "typical" distribution governing the observations in all the other coordinates. Nothing is known about the outlier and the typical distributions except that they are distinct and have full supports. The goal is to design a universal test to best discern the outlier coordinate. A universal test based on the generalized likelihood principle is proposed and is shown to be universally exponentially consistent, and a single-letter characterization of the error exponent achievable by the test is derived. It is shown that as the number of coordinates approaches infinity, our universal test is asymptotically efficient. Specifically, it achieves a limiting error exponent that is equal to the largest achievable error exponent when the outlier and typical distributions are both known.**

I. Introduction

We consider the following inference problem, which we term *outlier hypothesis testing*. In vector observations each with $M \geq 3$ coordinates, it is assumed that there is one outlier coordinate. Specifically, when the i-th coordinate is the outlier, the distribution governing the observations in that coordinate is assumed to be distinct from that governing the observations in all the other coordinates, which all come from the same "typical" distribution. The goal is to design a test to decide which coordinate is the outlier. We will be interested in the *universal* setting of this problem, where the test has to perform well without any prior knowledge of the outlier and typical distributions except that they must be different and have full supports.

It is to be noted that our problem of outlier hypothesis testing is distinct from that of statistical *outlier detection* [1], [2]. In outlier detection, the goal is to efficiently winnow out a few outlier observations from a single sequence of observations. The outlier observations are assumed to follow a different generating mechanism from that governing the normal observations. Statistical outlier detection is typically used to preprocess large data sets, to obtain clean data that is used for some purpose such as inference and control. The outlier hypothesis testing problem that we study here arises in event detection and environment monitoring in sensor networks [3], understanding of visual search in humans and animals [4], fraud and anomaly detection [5], [6] in large data sets, and optimal search and target tracking [7].

Universal outlier hypothesis testing is part of a broader class of *composite hypothesis testing* problems in which there is uncertainty in the probabilistic laws associated with some or all of the hypotheses. To solve these problems, a popular philosophy of test design is the *generalized likelihood principle* [8], [9]. The optimality of the generalized likelihood ratio test has been examined under various conditions [9], [10]. Universal outlier hypothesis testing is closely related to homogeneity testing and classification [11]–[15], both of which can be formulated as composite hypothesis testing problems. In homogeneity testing, one wishes to decide whether or not the two samples come from the same probabilistic law. In classification problems, a set of test data is classified based on another set of pre-acquired training data containing observations whose class membership is known. In [14], [15], a classifier based on the generalized likelihood principle was shown to be optimal under the asymptotic Neyman-Pearson criterion.

A metric that is commonly used to quantify the performance of a universal test is *consistency*. A universal test is *consistent* if the error probability approaches zero as the sample size goes to infinity, and is *exponentially consistent* if the decay is exponential with sample size. In our outlier hypothesis testing problem, we have neither any information regarding the outlier and the typical distributions, nor any training data to learn these distributions before the detection is performed. The only information we are given is that there is *exactly* one outlier coordinate in the observation vector, while the rest of the coordinates come from the same typical distribution. In other words, the only prior knowledge we have about the hypotheses is in terms of the structure of the joint distribution of the observation vector under each hypothesis. As a consequence, it is not clear at the outset that a universally exponentially consistent test should exist, and even if it does, it is not clear what its structure and performance should be.

Our technical contributions are as follows. First, we propose a universal test that follows the same principle as that underlying the generalized likelihood ratio test [8], [9]. When only the typical distribution is known, we show that our test achieves the same optimal error exponent as in the case where both the typical and outlier distributions are also known. We then consider the completely universal setting where both the typical and outlier distributions are unknown, and prove that our test is *universally exponentially consistent* for all $M \geq 3$. We also establish that as M goes to infinity, the error exponent achievable by our universal test converges to the optimal error exponent corresponding to the case where both the typical and outlier distributions are known.

II. Preliminaries

Throughout the paper, random variables are denoted by capital letters, and their realizations are denoted by the corresponding lower-case letters. All random variables are assumed to take values in finite sets, and all logarithms are the natural ones.

For a finite set $\mathcal{Y}$, let $\mathcal{Y}^m$ denote the m Cartesian product of $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y})$ denote the set of all probability mass functions (pmfs) on $\mathcal{Y}$. The empirical distribution of a sequence $\boldsymbol{y} = y^m = (y_1, \ldots, y_m) \in \mathcal{Y}^m$, denoted by $\gamma = \gamma_{\boldsymbol{y}} \in \mathcal{P}(\mathcal{Y})$, is defined as

$$\gamma(y) \triangleq \frac{1}{m} \big| \{k = 1, \ldots, m : y_k = y\} \big|,$$

$y \in \mathcal{Y}$.

Consider n independent and identically distributed (i.i.d.) vector observations, each of which has $M \geq 3$ independent coordinates. We denote the i-th coordinate of the k-th observation by $Y_k^{(i)} \in \mathcal{Y}$. It is assumed that only one coordinate is the "outlier," i.e., the observations in that coordinate are uniquely distributed (i.i.d.) according to the "outlier" distribution $\mu \in \mathcal{P}(\mathcal{Y})$, while all the other coordinates are commonly distributed according to the "typical" distribution $\pi \in \mathcal{P}(\mathcal{Y})$. *Nothing is known about μ and π except that $\mu \neq \pi$, and that each of them has full support.* Clearly, if $M = 2$, either coordinate can be considered as an outlier; hence, it becomes degenerate to consider outlier hypothesis testing in this case.

When the i-th coordinate is the outlier, the joint distribution of all the observations is

$$\begin{aligned} p_i\big(y^{Mn}\big) &= p_i\left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \\ &= \prod_{k=1}^{n} \Big\{ \mu\big(y_k^{(i)}\big) \prod_{j \neq i} \pi\big(y_k^{(j)}\big) \Big\}, \end{aligned}$$

where

$$\boldsymbol{y}^{(i)} = \left(y_1^{(i)}, \ldots, y_n^{(i)}\right), \; i = 1, \ldots, M.$$

The test for the outlier coordinate is done based on a *universal* rule $\delta : \mathcal{Y}^{Mn} \rightarrow \{1, \ldots, M\}$. In particular, the test δ is not allowed to depend on (μ, π).

For a universal test, the maximal error probability, which will be a function of the test and (μ, π), is

$$e\big(\delta, (\mu, \pi)\big) \triangleq \max_{i=1,\ldots,M} \sum_{y^{Mn}:\; \delta(y^{Mn}) \neq i} p_i\big(y^{Mn}\big),$$

and the corresponding error exponent is defined as

$$\alpha\big(\delta, (\mu, \pi)\big) \triangleq \lim_{n \to \infty} -\frac{1}{n} \log e\big(\delta, (\mu, \pi)\big).$$

The following technical facts will be useful; their derivations can be found in [16]. Consider random variables Y^n which are i.i.d. according to $p \in \mathcal{P}(\mathcal{Y})$. Let $y^n \in \mathcal{Y}^n$ be a sequence with an empirical distribution $\gamma \in \mathcal{P}(\mathcal{Y})$. It follows that the probability of such sequence y^n, under p and under the i.i.d. assumption, is

$$p(y^n) = \exp\Big\{ -n \big(D(\gamma \| p) + H(\gamma)\big) \Big\}, \qquad (1)$$

where $D(\gamma \| p)$ and $H(\gamma)$ are the relative entropy of γ and p, and entropy of γ, defined as

$$D(\gamma \| p) \triangleq \sum_{y \in \mathcal{Y}} \gamma(y) \log \frac{\gamma(y)}{p(y)},$$

and

$$H(\gamma) \triangleq -\sum_{y \in \mathcal{Y}} \gamma(y) \log \gamma(y),$$

respectively. Consequently, it holds that for each y^n, the pmf p that maximizes $p(y^n)$ is $p = \gamma$, and the associated maximal probability of y^n is

$$\gamma(y^n) = \exp\big\{ -nH(\gamma) \big\}. \qquad (2)$$

III. Proposed Universal Test

We now describe our universal test in two setups when only π is known, and when neither μ nor π is known, respectively.

For each $i = 1, \ldots, M$, denote the empirical distributions of $\boldsymbol{y}^{(i)}$ by γ_i. Note that the normalized log-likelihood of y^{Mn} when the i-th coordinate is the outlier is

$$L_i\left(y^{Mn}\right) \triangleq -\frac{1}{n} \log\left(p_i\left(y^{Mn}\right)\right),$$

for $i = 1, \ldots, M$. It now follows from (1) that

$$\begin{aligned} L_i\left(y^{Mn}\right) = &-\big[H\left(\gamma_i\right) + D\left(\gamma_i \| \mu\right)\big] \\ &- (M-1)\Big[H\Big(\tfrac{\sum_{j \neq i} \gamma_j}{M-1}\Big) + D\Big(\tfrac{\sum_{j \neq i} \gamma_j}{M-1} \Big\| \pi\Big)\Big], \end{aligned} \qquad (3)$$

and

$$L_i\left(y^{Mn}\right) = -\big[H(\gamma_i) + D(\gamma_i\|\mu)\big] - \sum_{j\neq i}\big[H(\gamma_j) + D(\gamma_j\|\pi)\big], \qquad (4)$$

for $i = 1, \ldots, M$.

When π is known and μ is unknown, we compute the generalized log-likelihood of y^{Mn} by replacing μ in (4) with its maximum likelihood (ML) estimate $\hat{\mu}_i \triangleq \gamma_i,\ i = 1, \ldots, M$, as

$$L_i^{\text{typ}}\big(y^{Mn}\big) \ = \ -H(\gamma_i) - \sum_{j\neq i}\big[H(\gamma_j) + D(\gamma_j\|\pi)\big]. \quad (5)$$

Similarly, when neither μ nor π is known, we compute the generalized log-likelihood of y^{Mn} by replacing the μ and π in (3) and (4) with their maximum likelihood (ML) estimates $\hat{\mu}_i \triangleq \gamma_i$, and $\hat{\pi}_i \triangleq \frac{\sum_{j\neq i}\gamma_j}{M-1},\ i = 1, \ldots, M$

$$L_i^{\text{univ}}\big(y^{Mn}\big) = \ -H(\gamma_i) - \sum_{j\neq i}\left[H(\gamma_j) + D\left(\gamma_j \middle\| \frac{\sum_{k\neq i}\gamma_k}{M-1}\right)\right]. \quad (6)$$

Finally, we decide upon the coordinate with the largest generalized log-likelihood to be the outlier. Using (5), (6), our universal tests in the two cases can be described respectively as

$$\delta\big(y^{Mn}\big) \ = \ \underset{i=1,\ldots,M}{\text{argmin}}\ U_i^{\text{typ}}\big(y^{Mn}\big), \qquad (7)$$

when only π is known, where for each $i = 1, \ldots, M$,

$$U_i^{\text{typ}}\big(y^{Mn}\big) \ \triangleq \ \sum_{j\neq i} D(\gamma_j\|\pi), \qquad (8)$$

and

$$\delta\big(y^{Mn}\big) \ = \ \underset{i=1,\ldots,M}{\text{argmax}}\ U_i^{\text{univ}}\big(y^{Mn}\big), \qquad (9)$$

when neither μ nor π is known, where for each $i = 1, \ldots, M$,

$$U_i^{\text{univ}}\big(y^{Mn}\big) \ \triangleq \ \sum_{j\neq i} D\left(\gamma_j \,\middle\|\, \frac{\sum_{k\neq i}\gamma_k}{M-1}\right). \qquad (10)$$

IV. Results

Our first theorem in this section characterizes the optimal exponent for the maximal error probability when both μ and π are known, and when only π is known.

Theorem 1. *For every $M \geq 3$, when μ and π are both known, the optimal exponent for the maximal error probability is equal to*

$$2B(\mu, \pi), \qquad (11)$$

where $B(\mu,\pi)$ is the Bhattacharyya distance between μ and π, $\mu \neq \pi$, which is defined as

$$B(\mu,\pi) \ \triangleq \ -\log\Big(\sum_{y\in\mathcal{Y}} \mu(y)^{\frac{1}{2}}\pi(y)^{\frac{1}{2}}\Big) \ > \ 0.$$

Furthermore, the error exponent in (11) is achievable by a test that uses only the knowledge of π. In particular, such a test is our proposed test in (7), (8).

Consequently, in the completely universal setting, when nothing is known about μ and π except that $\mu \neq \pi$, and both μ and π have full supports, it holds that for any universal test δ,

$$\alpha\big(\delta, (\mu, \pi)\big) \ \leq \ 2B(\mu, \pi). \qquad (12)$$

Notwithstanding the result in Theorem 1, without knowing either μ or π, it is not clear at the outset that we can design a universal test δ that yields $\alpha\left(\delta, (\mu, \pi)\right) > 0$ for *every* $\mu, \pi,\ \mu \neq \pi$. One of our main contributions in this paper is that *our proposed universal test in (9) and (10) is indeed universally exponentially consistent*. We also characterize the error exponent achievable by our proposed universal test.

Theorem 2. *Our proposed universal test δ in (9) and (10) is universally exponentially consistent. Furthermore, for every $\mu, \pi \in \mathcal{P},\ \mu \neq \pi$, it holds that*

$$\alpha\big(\delta, (\mu, \pi)\big) = \min_{q_1,\ldots,q_M} D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right) + \ldots + D\left(q_M\|\pi\right), \qquad (13)$$

where the minimum above is over the set of $(q_1, \ldots, q_M)$ such that

$$\sum_{j\neq 1} D\left(q_j \,\middle\|\, \frac{\sum_{k\neq 1} q_k}{M-1}\right) \geq \sum_{j\neq 2} D\left(q_j \,\middle\|\, \frac{\sum_{k\neq 2} q_k}{M-1}\right). \quad (14)$$

Note that for any fixed $M \geq 3$, $\epsilon > 0$, regardless of which coordinate is the outlier, it holds that the random empirical distributions $(\gamma_1, \ldots, \gamma_M)$ satisfy

$$\lim_{n\rightarrow\infty} \mathbb{P}_i\left\{\left\| \tfrac{1}{M}\textstyle\sum_{j=1}^{M}\gamma_j - \left(\tfrac{1}{M}\mu + \tfrac{M-1}{M}\pi\right)\right\|_1 > \epsilon\right\} = 0, \qquad (15)$$

where $\|\cdot\|_1$ denotes the 1-norm of the argument distribution. Since $\frac{1}{M}\mu + \frac{M-1}{M}\pi \rightarrow \pi$ as $M \rightarrow \infty$, heuristically speaking, a consistent estimate of the typical distribution can readily be obtained asymptotically in M at the outset from the entire observations before deciding upon which coordinate is the outlier. This observation and the second assertion of Theorem 1 motivate our study of the asymptotic performance of our proposed universal test in (9), (10) when $M \rightarrow \infty$.

Our last result in this section shows that in the completely universal setting, as $M \rightarrow \infty$, our proposed universal test in (9), (10) achieves the optimal error exponent in (11) corresponding to the case in which *both*

μ and π are known.

Theorem 3. *For each $M \geq 3$, the exponent for the maximal error probability achievable by our proposed universal test δ in (9), (10) is lower bounded by*

$$\min_{\substack{q \,\in\, \mathcal{P}(\mathcal{Y}) \\ D(q\|\pi) \,\leq\, \frac{1}{M-1}\left(2B(\mu,\pi)+C_\pi\right)}} \quad 2\,B(\mu\,,\,q)\,, \tag{16}$$

where $C_\pi \triangleq -\log\left(\min_{y\in\mathcal{Y}} \pi(y)\right) < \infty$ by the fact that π has a full support.

The lower bound for the error exponent in (16) is nondecreasing in $M \geq 3$. Furthermore, as $M \to \infty$, this lower bound converges to the optimal error exponent $2B(\mu, \pi)$; hence, it holds that

$$\lim_{M\to\infty} \alpha\big(\delta, (\mu, \pi)\big) \;=\; 2B(\mu, \pi). \tag{17}$$

V. Discussion

The appealing properties of our universal test, i.e., the universally exponential consistency and the asymptotic efficiency, can also be shown to hold in the case with more than one outlier coordinate as long as the number of outliers is fixed and known [17].

It is to be noted that our results rely critically on the assumption that the number of outlier coordinates is *known exactly*. For example, in the case with only one outlier, if only one additional hypothesis corresponding to the situation with no outlier is present, then the nature of the problem changes completely. In particular, for this new setup, it can be shown that there cannot exist any universally exponentially consistent test *even when the typical distribution is known.*

We end with a discussion of possible extensions of our results. First, it is worth noting that although efficient in many cases, generalized likelihood tests fall short of optimality in some situations [18], [19]. A different approach, namely, the "competitive minimax" approach, proposed by Feder and Merhav, is aimed at minimizing the worst-case ratio between the probability of error of a universal test and the minimum probability of error when the underlying distributions are fully known [18]. Under such competitive minimax performance criteria, it is interesting to see what the structure and performance of an optimal test are in the universal outlier hypothesis testing problem. Another interesting way to extend the results of this paper would be to consider models with the size of the alphabet being large compared to the number of samples from each coordinate [19], [20].

VI. Sketches of Proofs

Our proofs rely on the following two lemmas, and the first of which is an extension of Sanov's theorem.

Lemma 1. *Let $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(J)}$ be mutually independent random vectors with each $\boldsymbol{Y}^{(j)}$, $j = 1, \ldots, J$, being n i.i.d. repetitions of a random variable distributed according to $p_j \in \mathcal{P}(\mathcal{Y})$ with a full support. Let A_n be the set of all J tuples $(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(J)}) \in \mathcal{Y}^{Jn}$ whose empirical distributions $(\gamma_1, \ldots, \gamma_J) = (\gamma_{\boldsymbol{y}^{(1)}}, \ldots, \gamma_{\boldsymbol{y}^{(J)}})$ lie in a closed set $E \in \mathcal{P}(\mathcal{Y})^J$. Then, it holds that*

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\Big\{\Big(\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(J)}\Big) \in A_n\Big\} \;=\; \min_{(q_1,\ldots,q_J)\,\in\, E} \;\sum_{j=1}^{J} D(q_j\|p_j).$$

Lemma 2. *For any two pmfs p_1, $p_2 \in \mathcal{P}(\mathcal{Y})$ with full supports, it holds that*

$$2B\,(p_1\,,\,p_2) \;=\; \min_{q\,\in\,\mathcal{P}(\mathcal{Y})} D\,(q\|p_1) + D\,(q\|p_2)\,. \tag{18}$$

In particular, the minimum on the right side of (18) is achieved by

$$q^\star \;=\; \frac{p_1^{\frac{1}{2}}(y)p_2^{\frac{1}{2}}(y)}{\sum\limits_{y\in\mathcal{Y}} p_1^{\frac{1}{2}}(y)p_2^{\frac{1}{2}}(y)},\; y \in \mathcal{Y}. \tag{19}$$

A. Sketch of Proof of Theorem 1

When μ and π are known, it is clear that the test which maximizes the error exponent is the ML one. In particular, for any $y^{Mn} = \big(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\big) \in \mathcal{Y}^{Mn}$, with $\gamma_{\boldsymbol{y}^{(i)}} = \gamma_i$, $i = 1, \ldots, M$, the ML test is

$$\delta(y^{Mn}) \;=\; \operatorname*{argmin}_{i=1,\ldots,M} \; U_i(y^{Mn}),$$

where for each $i = 1, \ldots, M$,

$$U_i(y^{Mn}) \;\triangleq\; D\,(\gamma_i\|\mu) + \sum_{j\neq i} D\,(\gamma_j\|\pi)\,. \tag{20}$$

By the symmetry of the problem, the optimal error exponent is the same as the exponent of the following event

$$\mathbb{P}_1\{U_1 \geq U_2\} \;=\; \mathbb{P}_1\{D\,(\gamma_1\|\mu) + D\,(\gamma_2\|\pi) \geq D\,(\gamma_1\|\pi) + D\,(\gamma_2\|\mu)\}.$$

Applying Lemma 1 with $J = 2$, $p_1 = \mu$, $p_2 = \pi$,

$$E = \left\{ q_1, q_2 \;:\; \begin{array}{c} D\,(q_1\|\mu) + D\,(q_2\|\pi) \\ \geq\; D\,(q_1\|\pi) + D\,(q_2\|\mu) \end{array} \right\},$$

we get that the exponent for $\mathbb{P}_1\{U_1 \geq U_2\}$ is given by the value of the following optimization problem

$$\min_{\substack{q_1, q_2 \in \mathcal{P}(\mathcal{Y}) \\ D(q_1\|\mu)+D(q_2\|\pi)\geq D(q_1\|\pi)+D(q_2\|\mu)}} \quad D\,(q_1\|\mu) + D\,(q_2\|\pi)\,. \tag{21}$$

The optimization problem (21) is convex and its solution can be easily computed to be $2B(\mu, \pi)$.

When only π is known, it follows from the same argument leading to (21) that the achievable error exponent of our proposed test δ' in (7), (8) is given by

$$\min_{\substack{q_1, q_2 \in \mathcal{P}(\mathcal{Y}) \\ D(q_2 \| \pi) \ \geq \ D(q_1 \| \pi)}} \quad D\left(q_1 \| \mu\right) + D\left(q_2 \| \pi\right). \tag{22}$$

The optimal value of (22) can be shown to be $2B(\mu, \pi)$.

B. Sketch of Proof of Theorem 2

When μ and π are unknown, we adopt our universal test in (9) and (10). Applying Lemma 1, it follows from the same argument leading to (21) that the achievable error exponent of our universal test is given by the optimal value of (13).

Although a closed-form solution to (13) is not available, we show that the value of (13) is strictly positive for every $\mu, \pi, \mu \neq \pi$. In particular, the objective function is continuous in $q_1, \ldots, q_M$ and the constraint in (14) is compact. The claim then follows from the fact that the value of the objective function in (13) is strictly positive at every feasible $q_1, \ldots, q_M$. Thus, our proposed test is indeed universally exponentially consistent.

C. Sketch of Proof of Theorem 3

By the continuity of the objective function on the right-side of (13) and the compactness of the constraint set (14), for each $M \geq 3$, the optimal value on the right-side of (13), denoted by $V^\star$, is achieved by some $(q_1^\star, \ldots, q_M^\star)$. It follows from (13), (14) and Lemma 2 that

$$\begin{aligned} V^\star \ &\geq \ D(q_1^\star \,\|\, \mu) + \sum_{j \neq 1} D\left(q_j^\star \| \pi\right) \\ &\geq \ D(q_1^\star \,\|\, \mu) + D\Big(q_1^\star \,\Big\|\, \tfrac{\sum_{k \neq 2} q_k^\star}{M-1}\Big) \\ &\geq \ 2B\Big(\mu, \tfrac{\sum_{k \neq 2} q_k^\star}{M-1}\Big). \end{aligned} \tag{23}$$

On the other hand, it follows from (12) that the value on the right-side of (13), $V^\star$, satisfies

$$\begin{aligned} 2B(\mu, \pi) \ &\geq \ \sum_{j=3}^{M} D\left(q_j^\star \,\|\, \pi\right) \\ &\geq \ (M-2)\, D\Big(\tfrac{1}{M-2} \textstyle\sum_{k=3}^{M} q_k^\star \,\Big\|\, \pi\Big). \end{aligned} \tag{24}$$

Combining (23) and (24), we get that $V^\star$ is lower bounded by

$$\min_{\substack{q \in \mathcal{P}(\mathcal{Y}) \\ D(q \| \pi) \leq \frac{1}{M-1}(2B(\mu,\pi) + C_\pi)}} \quad 2\, B\left(\mu\,,\, q\right), \tag{25}$$

where $C_\pi \triangleq -\log\big(\min_{y \in \mathcal{Y}} \pi(y)\big)$.

The assertion in (17) follows by virtue of fact that for any $\mu, \pi \in \mathcal{P}(\mathcal{Y})$ with full supports, it holds that

$$\lim_{M \to \infty} \frac{1}{M-1}\big(2B(\mu, \pi) + C_\pi\big) = 0.$$

VII. Acknowledgements

This work was supported by the Air Force Office of Scientific Research (AFOSR) under the Grant FA9550-10-1-0458 and by the National Science Foundation under Grant NSF CCF 11-11342, through the University of Illinois at Urbana-Champaign.

References

[1] V. Barnett, "The study of outliers: purpose and model," *Appl. Stat.*, vol. 27, no. 3, pp. 242–250, 1978.

[2] D. Hawkins, *Identification of Outliers*. Chapman and Hall, 1980.

[3] J. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, pp. 16–25, 2007.

[4] N. K. Vaidhiyan, S. P. Arun and R. Sundaresan, "Active sequential hypothesis testing with application to a visual search problem," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 2201–2205.

[5] R. J. Bolten and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, pp. 235–249, 2002.

[6] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15.1–15.58, 2009.

[7] L. D. Stone, *Theory of Optimal Search*. Topics in Operations Research Series, INFORMS, 2004.

[8] V. H. Poor, *An Introduction to Signal Detect and Estimation*. Springer, 1994.

[9] O. Zeitouni, J. Ziv and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inf. Theory*, vol. 38, pp. 1597–1602, 1992.

[10] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–401, 1965.

[11] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, pp. 250–254, 1911.

[12] O. Shiyevitz, "On Rényi measures and hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 31-Aug. 5 2011, pp. 894–898.

[13] J. Unnikrishnan, "On optimal two sample homogeneity tests for finite alphabets," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 1-6 2012, pp. 2027–2031.

[14] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. 34, pp. 278–286, 1988.

[15] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, pp. 401–408, 1989.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, Inc., 1991.

[17] Y. Li, S. Nitinawarat and V. V. Veeravalli, "Universal outlier hypothesis testing," http://arxiv.org/abs/1302.4776, submitted to *IEEE Tran. Inf. Theory*, April, 2013.

[18] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1504–1517, 2002.

[19] B. G. Kelly, A. B. Wagner, T. Tularak and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE Trans. Inf. Theory*, vol. 59, pp. 782–795, 2013.

[20] D. Huang and S. P. Meyn, "Classification with high-dimensional sparse samples," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 2586–2590.