Universal Outlier Hypothesis Testing via Mean- and Median-Based Tests

Bernhard C. Geiger*†¶, Tobias Koch‡§, Josipa Mihaljević†, and Maximilian Toller†
*Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria
†Know Center Research GmbH, 8010 Graz, Austria
¶Graz Center for Machine Learning, 8010 Graz, Austria
‡Signal Theory and Communications Department, Universidad Carlos III de Madrid, 28911 Leganés, Spain
§Gregorio Marañón Health Research Institute, 28007 Madrid, Spain
e-mails: geiger@ieee.org, {jmihaljevic,mtoller}@know-center.at, tkoch@ing.uc3m.es

***Abstract*—Universal outlier hypothesis testing refers to a hypothesis testing problem where one observes a large number of length-n sequences—the majority of which are distributed according to the typical distribution π and a small number are distributed according to the outlier distribution μ—and one wishes to decide, which of these sequences are outliers without having knowledge of π and μ. In contrast to previous works, in this paper it is assumed that both the number of observation sequences and the number of outlier sequences grow with the sequence length. In this case, the typical distribution π can be estimated by computing the mean over all observation sequences, provided that the number of outlier sequences is sublinear in the total number of sequences. It is demonstrated that, in this case, one can achieve the error exponent of the maximum likelihood test that has access to both π and μ. However, this mean-based test performs poorly when the number of outlier sequences is proportional to the total number of sequences. For this case, a median-based test is proposed that estimates π as the median of all observation sequences. It is demonstrated that the median-based test achieves again the error exponent of the maximum likelihood test that has access to both π and μ, but only with probability approaching one. To formalize this case, the typical error exponent—similar to the typical random coding exponent introduced in the context of random coding for channel coding—is proposed.**

I. Introduction

The problem of *outlier hypothesis testing*, introduced by Li, Nitinawarat, and Veeravalli [1], consists in deciding among a large number of sequences which ones are outliers. More precisely, in outlier hypothesis testing, we observe M independent length-n sequences $\mathbf{Y}_1, \dots, \mathbf{Y}_{\mathsf{M}}$, $\mathsf{M}-\mathsf{T}$ of which are independent and identically distributed (i.i.d.) according to the *typical distribution* π, and $\mathsf{T} < \mathsf{M}/2$ are i.i.d. according to the *outlier distribution* $\mu \neq \pi$, both taking values from the same finite alphabet $\mathcal{Y}$. The goal is to decide which sequences are distributed according to μ, i.e., to identify the outlier sequences. For any fixed M, this can be formulated as the problem of finding optimizers of a functional over the search space $[\mathsf{M}] \triangleq \{1, \dots, \mathsf{M}\}$.

We shall describe the outlier hypothesis test by the function $\delta\colon \mathcal{Y}^{\mathsf{M}n} \to \bar{\mathcal{S}}$, where $\bar{\mathcal{S}}$ denotes the set of size-T subsets of $[\mathsf{M}]$. In words, the function $\delta(\cdot)$ takes the M sequences $\mathbf{Y}_1, \dots, \mathbf{Y}_{\mathsf{M}}$ as input and produces the set of indices $\mathcal{S}$ indicating the outlier sequences. The performance of this test is measured by the maximal error probability

$$\epsilon(\delta) = \max_{\mathcal{S}\in\bar{\mathcal{S}}} \sum_{\mathbf{y}_1^{\mathsf{M}}\colon \delta(\mathbf{y}_1^{\mathsf{M}})\neq\mathcal{S}} P_{\mathcal{S}}(\mathbf{y}_1^{\mathsf{M}}) \tag{1}$$

where $P_{\mathcal{S}}$ denotes the joint distribution of $\mathbf{Y}_1, \dots, \mathbf{Y}_{\mathsf{M}}$ when the outlier sequences have indices $\mathcal{S}$, and where we use the notation $\mathbf{y}_1^{\mathsf{M}}$ to denote the sequence $\mathbf{y}_1, \dots, \mathbf{y}_{\mathsf{M}}$. More precisely, we study the decay at which $\epsilon(\delta)$ tends to zero as $n \to \infty$ by considering the corresponding error exponent

$$\alpha(\delta) \triangleq \varlimsup_{n\to\infty} -\frac{1}{n}\log\epsilon(\delta) \tag{2}$$

where $\log(\cdot)$ is the binary logarithm function and $\varlimsup$ denotes the *limit superior*.

If both distributions π and μ, and the number T of outlier sequences are known, then the error exponent of the maximum likelihood test is $2B(\pi,\mu)$ [1, Prop. 7], where

$$B(\pi,\mu) \triangleq -\log\left(\sum_{y\in\mathcal{Y}} \pi(y)^{\frac{1}{2}}\mu(y)^{\frac{1}{2}}\right) \tag{3}$$

is the Bhattacharyya distance between π and μ. The same exponent can also be achieved by a test that has only access to π but is ignorant of μ [1, Th. 8].

In the universal setting, where neither π nor μ are known, the maximum likelihood test is inapplicable. Instead, Li, Nitinawarat, and Veeravalli consider the generalized likelihood ratio test (GLRT), which is given by [1, eq. (37)]

$$\begin{aligned}&\delta_{\mathrm{Li}}(\mathbf{Y}_1, \dots, \mathbf{Y}_{\mathsf{M}})\\ &\quad= \operatorname*{argmin}_{\mathcal{S}\in\bar{\mathcal{S}}} \sum_{j\notin\mathcal{S}} D\left(\mathsf{P}_{\mathbf{Y}_j}\middle\|\frac{\sum_{k\notin\mathcal{S}}\mathsf{P}_{\mathbf{Y}_k}}{\mathsf{M}-\mathsf{T}}\right)\end{aligned} \tag{4}$$

where

$$\mathsf{P}_{\mathbf{Y}_k}(y) \triangleq \frac{1}{n}\sum_{\ell=1}^{n} \mathbf{1}\left\{Y_{k,\ell} = y\right\}, \quad y\in\mathcal{Y} \tag{5}$$

denotes the empirical distribution (type) of $\mathbf{Y}_k = (Y_{k,1}, \dots, Y_{k,n})$, and $\mathbf{1}\{\cdot\}$ denotes the indicator

Author names are listed in alphabetical order.

function. For the GLRT, they derive a lower bound on the error exponent, which is strictly smaller than $2B(\pi,\mu)$, but converges to $2B(\pi,\mu)$ as the number of observation sequences M tends to infinity [1, Th. 10]. This agrees with the intuition that, as the number of observation sequences tends to infinity, we can accurately estimate the typical distribution by averaging over all observations.

The work of Li, Nitinawarat, and Veeravalli [1] has subsequently been extended in various directions. For example, a test for $\mathsf{T} \leq 1$ with rejection option, i.e., where the test can decide that none of the observation sequences is an outlier, was investigated in [2]. Nitinawarat and Veeravalli [3] considered the case where in exactly one sequence a change point occurs at time $1 \leq \lambda \leq n$ and proposed a test for identifying the sequence and λ. Bu *et al.* [4] proposed tests for continuous distributions π and μ. Sequential detection, i.e., early stopping, was considered in [5]–[7]. Acharya *et al.* [8] studied the setting where the sample size n is smaller than the alphabet size $|\mathcal{Y}|$ of the discrete distributions.

The aforementioned works all concern the case where the sample size n is significantly larger than the number M of observation sequences. However, sometimes the opposite case is also of interest. A practical example are dense sensor networks, where each of the M sensors can only sense over short intervals or with low sampling rates due to energy limitations. Another example occurs in the medical domain, where a large number of patients are recorded for a comparably short time, such as for ECG/EEG data or voice recordings. A large number of observation sequences has the advantage that the typical distribution can be estimated accurately as

$$\hat{\pi}(y) = \frac{1}{\mathsf{M}} \sum_{k=1}^{\mathsf{M}} \mathsf{P}_{\mathbf{Y}_k}(y), \quad y \in \mathcal{Y} \tag{6}$$

provided that the number T of outlier sequences is sublinear in M. On the negative side, the GLRT may become infeasible if the number of observation sequences is large. Indeed, the test searches over all size-T subsets of $[\mathsf{M}]$ as candidate outlier sets. It thus has a computational complexity of $\mathcal{O}(\mathsf{M}^{\mathsf{T}})$, which becomes prohibitive if M and T are large. Mathematically, we model the case of a large number of observation sequences by making the parameters M (number of observation sequences) and T (number of outlier sequences) dependent on the sequence length n. As a consequence, the approach followed by Li, Nitinawarat, and Veeravalli [1] of applying Sanov's theorem to express the error exponent as a minimization of relative entropies and then bounding this minimum cannot be followed for the GLRT, since the subexponential terms of the error probability depend exponentially on the alphabet size $\mathcal{Y}^{\mathsf{M}}$ of the sequences $\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}}$ and may become significant for an n-dependent M.

To sidestep this problem, we propose the *mean-based test*

$$\delta_{\mathrm{mean}}(\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}}) = \operatorname*{argmax}_{\mathcal{S}\in\bar{\mathcal{S}}} \sum_{j\in\mathcal{S}} D\left(\mathsf{P}_{\mathbf{Y}_j} \middle\| \hat{\pi}\right) \tag{7}$$

whose error probability, conditioned on $\hat{\pi}$, can be analyzed analytically. In particular, we show that, if $\mathsf{T} = o(\mathsf{M})$ and M grows superlinearly but subexponentially in n, then the mean-based test achieves the optimal error exponent $2B(\pi,\mu)$.

If the outlier sequences represent a positive fraction of all sequences, i.e., if $\mathsf{T} = c\mathsf{M}$, $0 < c < \frac{1}{2}$, then the mean over all types is biased, which negatively affects the error exponent of the mean-based test (7). In this case, the median over all types

$$\bar{\pi}(y) = \frac{\operatorname{median}\{\mathsf{P}_{\mathbf{Y}_1}(y),\dots,\mathsf{P}_{\mathbf{Y}_{\mathsf{M}}}(y)\}}{\sum_{y'\in\mathcal{Y}} \operatorname{median}\{\mathsf{P}_{\mathbf{Y}_1}(y'),\dots,\mathsf{P}_{\mathbf{Y}_{\mathsf{M}}}(y')\}}, \; y \in \mathcal{Y} \tag{8}$$

may provide a better estimate of π, since it is more robust against outliers. As we shall show, the median becomes accurate as the sequence length n tends to infinity but, in contrast to the mean, its accuracy does not improve with M. This implies that the error exponent of a median-based test that replaces in (7) the mean-based estimate $\hat{\pi}$ by the median-based estimate $\bar{\pi}$ is poor, despite the robustness of the median against outliers, and despite numerical results that seem to suggest otherwise. We thus propose a two-step median-based test that computes the median-based estimate $\bar{\pi}$ from a part of the sequences $\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}}$ and uses the remaining part for outlier testing. As a less pessimistic performance measure, we then propose the *typical* error exponent, defined as the average error exponent averaged over all realizations of $\bar{\pi}$. The typical error exponent is similar to the typical random coding exponent introduced in the context of random coding for channel coding; see, e.g., [9]–[12]. We show that the typical error exponent of the two-step median-based test is equal to the optimal error exponent $2B(\pi,\mu)$. Thus, the test achieves the optimal exponent with probability approaching one.

In terms of complexity, the mean-based test and the two-step median-based test have a computational complexity of $\mathcal{O}(\mathsf{M}\log\mathsf{M})$, while searching for a size-$c\mathsf{M}$ subset, as in the GLRT (4), has complexity $\mathcal{O}(2^{2\mathsf{M}}/\sqrt{\mathsf{M}})$.

The rest of this paper is organized as follows. Section II introduces the considered setup. Section III discusses the error exponent of the mean-based test. Section IV studies the two-step median-based test and presents its typical error exponent. Section V compares the performances of the proposed tests with that of the GLRT by means of numerical examples. Section VI concludes the paper with a discussion of the computation complexity of the considered tests. The full proofs of our results can be found in [13].

II. Setup

Suppose we observe M_n independent length-n sequences $\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}_n}$, $\mathsf{M}_n - \mathsf{T}_n$ of which are i.i.d. according to the typical distribution π and $\mathsf{T}_n < \mathsf{M}_n/2$ are i.i.d. according to the outlier distribution $\mu \neq \pi$. Here and throughout the rest of this paper, we add the subscript "n" to M and T to reflect in the notation that these two parameters may depend on the sequence length n. We assume that μ and π have the same alphabet $\mathcal{Y}$, which is assumed to be finite. We further assume that

$$\pi_{\min} \triangleq \min_{y\in\mathcal{Y}} \pi(y) > 0. \tag{9}$$

By the symmetry of the problem, we can assume without loss of generality that the first T_n sequences are outliers. Under

this assumption, the sequences $\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathsf{M}_n}$ have the joint distribution

$$P_{[\mathsf{T}_n]}(\mathbf{y}_1, \ldots, \mathbf{y}_{\mathsf{M}_n}) = \prod_{k=1}^{n} \prod_{i=1}^{\mathsf{T}_n} \mu(y_{i,k}) \prod_{j=\mathsf{T}_n+1}^{\mathsf{M}_n} \pi(y_{j,k}) \quad (10)$$

where $\mathbf{y}_j = (y_{j,1}, \ldots, y_{j,n})^T$, $j = 1, \ldots, \mathsf{M}_n$. We shall assume that the number of outlier sequences T_n is known. Any outlier test can thus be written as a function $\delta(\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathsf{M}_n})\colon \mathcal{Y}^{n\mathsf{M}_n} \to \bar{\mathcal{S}}$, where $\bar{\mathcal{S}}$ denotes the set of size-T_n subsets of $[\mathsf{M}_n]$, i.e., the set of all sets of cardinality T_n in the powerset $2^{[\mathsf{M}_n]}$. The maximal error probability can then be written as

$$\epsilon(\delta) = \sum_{\mathbf{y}_1^{\mathsf{M}_n}\colon \delta(\mathbf{y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n]} P_{[\mathsf{T}_n]}(\mathbf{y}_1^{\mathsf{M}_n}) \quad (11)$$

and the corresponding error exponent is as in (2).

For the two-step median-based test, we further introduce the *typical error exponent*. Specifically, suppose that we use the first $\lceil \rho n \rceil$, $\rho \in (0,1)$ samples of each one of the vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathsf{M}_n}$, denoted by $\mathbf{Y}_1^{(1)}, \ldots, \mathbf{Y}_{\mathsf{M}_n}^{(1)}$, to produce the median-based estimate $\bar{\pi}$ of π (cf. (8)) and the remaining samples $\mathbf{Y}_1^{(2)}, \ldots, \mathbf{Y}_{\mathsf{M}_n}^{(2)}$ to decide which sequences are outliers. Thus, in the second step, the outlier test is given by the function

$$\delta_{\text{median}}(\mathbf{Y}_1^{(2)}, \ldots, \mathbf{Y}_{\mathsf{M}_n}^{(2)} | \bar{\pi}) = \operatorname*{argmax}_{\mathcal{S} \in \bar{\mathcal{S}}} \sum_{j \in \mathcal{S}} D\left(\mathsf{P}_{\mathbf{Y}_j^{(2)}} \middle\| \bar{\pi}\right). \quad (12)$$

The maximal error probability in the second step can be written as

$$\epsilon(\delta_{\text{median}} | \bar{\pi}) = \sum_{\mathbf{y}_1^{\mathsf{M}_n}\colon \delta_{\text{median}}(\mathbf{y}_1^{\mathsf{M}_n} | \bar{\pi}) \neq [\mathsf{T}_n]} P^{(2)}_{[\mathsf{T}_n]}(\mathbf{y}_1^{\mathsf{M}_n}) \quad (13)$$

where $P^{(2)}_{[\mathsf{T}_n]}$ is as in (10) but with the index k going from $\lceil \rho n \rceil + 1$ to n. The typical error exponent is then defined as

$$\bar{\alpha}(\delta_{\text{median}}) \triangleq \varlimsup_{n \to \infty} -\frac{1}{n} \mathsf{E}\left[\log\big(\epsilon(\delta_{\text{median}} | \bar{\pi})\big)\right] \quad (14)$$

where the expectation is over all observation sequences $\mathbf{Y}_1^{(1)}, \ldots, \mathbf{Y}_{\mathsf{M}_n}^{(1)}$ used in step 1 to produce $\bar{\pi}$.

Jensen's inequality implies that the typical error exponent $\bar{\alpha}(\delta_{\text{median}})$ is never smaller than the error exponent $\alpha(\delta_{\text{median}})$. Intuitively, to obtain the optimal typical error exponent, it suffices that the estimate $\bar{\pi}$ is accurate with probability approaching one. In contrast, the error exponent requires the probability that $\bar{\pi}$ deviates from π to decay exponentially in n with a sufficiently large error exponent. Our numerical results in Section V suggest that such a requirement may be too pessimistic.

III. Mean-Based Test

We start by investigating the mean-based test in (7). The following theorem characterizes the performance of this test when the number of outlier sequences is sublinear in M_n.

Theorem 1: Assume the number of sequences M_n satisfies $\lim_{n\to\infty} \mathsf{M}_n/n = \infty$ and $\lim_{n\to\infty} \log(\mathsf{M}_n)/n = 0$. Further assume that the number of outlier sequences is known and satisfies $\mathsf{T}_n = o(\mathsf{M}_n)$. Then, for every pair of distributions $\mu \neq \pi$, the mean-based test (7) has the error exponent

$$\alpha_{\text{mean}} = 2B(\pi, \mu). \quad (15)$$

Proof (Sketch): Since the error exponent in the universal setting cannot be larger than the error exponent $2B(\pi,\mu)$ in the setting where π and μ are known, it suffices to bound the error exponent from below.

We next note that, by the assumption that $\mathsf{T}_n = o(\mathsf{M}_n)$, the estimate $\hat{\pi}$ converges to π as $\mathsf{M}_n \to \infty$. In fact, Hoeffding's inequality [14] implies that

$$\text{Prob}\big(\hat{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)\big) \leq 2|\mathcal{Y}| \exp\left(-2\frac{(\varepsilon \mathsf{M}_n - \mathsf{T}_n)^2}{\mathsf{M}_n}\right) \quad (16)$$

for every $\varepsilon > 0$, where

$$\mathcal{B}_{\varepsilon,\infty}(\pi) \triangleq \{Q \in \mathcal{P}(\mathcal{Y})\colon Q(y) \in [\pi(y) - \varepsilon, \pi(y) + \varepsilon], y \in \mathcal{Y}\}. \quad (17)$$

We can then upper-bound the error probability as

$$\epsilon(\delta_{\text{mean}}) \leq \text{Prob}\left(\delta_{\text{mean}}(\mathbf{Y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n], \hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right) + 2|\mathcal{Y}| \exp\left(-2\frac{(\varepsilon \mathsf{M}_n - \mathsf{T}_n)^2}{\mathsf{M}_n}\right) \quad (18)$$

whose error exponent is determined by the error exponent of the first term since, by the assumption that M_n is superlinear in n and that $\mathsf{T}_n = o(\mathsf{M}_n)$, we have that $\text{Prob}\big(\hat{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)\big)$ decays superexponentially in n. Furthermore, for discrete distributions, relative entropy is continuous. It thus follows that, for $\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$, $U_i(\hat{\pi}) \triangleq D(\mathsf{P}_{\mathbf{Y}_i} \| \hat{\pi})$ lies in $[U_i(\pi) - \varepsilon', U_i(\pi) + \varepsilon']$ for some ε' that vanishes as $\varepsilon \to 0$.

By assumption, the first T_n sequences are outliers. The first term on the RHS of (18) can thus be bounded as

$$\begin{aligned} &\text{Prob}\left(\delta_{\text{mean}}(\mathbf{Y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n], \hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right) \\ &\quad \leq P_{[\mathsf{T}_n]}\left(\min_{i \leq \mathsf{T}_n} U_i(\pi) \leq \max_{j > \mathsf{T}_n} U_j(\pi) + 2\varepsilon'\right) \\ &\quad \leq \mathsf{T}_n(\mathsf{M}_n - \mathsf{T}_n) P_{[\mathsf{T}_n]}\left(U_1(\pi) \leq U_{\mathsf{T}_n+1}(\pi) + 2\varepsilon'\right) \end{aligned} \quad (19)$$

where the first step follows because, for $\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$, we can approximate $U_i(\hat{\pi})$ by $U_i(\pi) \pm \varepsilon'$, and by then removing the condition $\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$; and the second step follows by writing the condition in the probability $P_{[\mathsf{T}_n]}(\cdot)$ as a union of events, and by applying the union bound.

Since $\mathsf{T}_n < \mathsf{M}_n/2$ and, by the theorem's assumption, M_n is subexponential in n, it follows that the error exponent of the RHS of (19) is determined by the exponent of $P_{[\mathsf{T}_n]}\left(U_1(\pi) \leq U_{\mathsf{T}_n+1}(\pi) + 2\varepsilon'\right)$. This exponent, in turn, can be computed by following the steps in [1] for the case where π is known. In particular, it was shown in [1, Th. 1] that the exponent of the probability $P_{[\mathsf{T}_n]}\left(U_1(\pi) \leq U_{\mathsf{T}_n+1}(\pi)\right)$ is equal to $2B(\pi,\mu)$. A simple continuity argument then yields that the exponent of $P_{[\mathsf{T}_n]}\left(U_1(\pi) \leq U_{\mathsf{T}_n+1}(\pi) + 2\varepsilon'\right)$ converges to $2B(\pi,\mu)$ as we let ε' tend to zero from above.

For the full proof, see [13, App. A]. ∎

Theorem 1 hinges on the assumption that T_n is sublinear in M_n, so the estimate $\hat{\pi}$ of π becomes accurate as $n \to \infty$. When T_n is proportional to M_n, this is no longer the case. In particular, if $\mathsf{T}_n = c\mathsf{M}_n$, $c > 0$, then $\hat{\pi}$ converges to the distribution $\nu = (1-c)\pi + c\mu$. By following the same steps as in the full proof of Theorem 1 in [13], it can be shown that the error exponent is lower-bounded as

$$\alpha_{\text{mean}} \geq \min_{Q_1, Q_2} \{D(Q_1\|\mu) + D(Q_2\|\pi)\} \tag{20}$$

where the minimum is over all distributions (Q_1, Q_2) satisfying $D(Q_2|\nu) - D(Q_1\|\nu) \geq 0$, which is smaller than $2B(\pi, \mu)$, cf. Fig. 2. Thus, when the number of outlier sequences is proportional to the total number of sequences M_n, the mean-based test does not achieve the exponent of the maximum-likelihood test when π and μ are known.

IV. Median-Based Test

Since the mean-based test performs subpar for $\mathsf{T}_n = c\mathsf{M}_n$, we consider in this section the median-based estimate (8), which is inherently robust against outliers as long as $c < \frac{1}{2}$. However, in contrast to the mean-based estimate $\hat{\pi}$, the convergence of the median-based estimate $\bar{\pi}$ to π is determined by the growth of n and not of M_n. Consequently, the probability that $\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)$ (with $\mathcal{B}_{\varepsilon,\infty}(\pi)$ defined in (17)) does not decay superexponentially in n, even if M_n is superlinear in n. Specifically, this probability can be bounded as follows:

Lemma 1: Consider the median-based estimate $\bar{\pi}$ of π, in (8), and assume that $\mathsf{T}_n < \mathsf{M}_n/2$. Then, for every $\varepsilon > 0$,

$$P_{[\mathsf{T}_n]}\big(\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)\big) \leq 2|\mathcal{Y}|(\mathsf{M}_n - \mathsf{T}_n)\exp(-2n\varepsilon'^2) \tag{21}$$

where $\varepsilon' = \varepsilon/(1+|\mathcal{Y}|)$.

Proof (Sketch): Let

$$m(y) \triangleq \text{median}\{\mathsf{P}_{\mathbf{Y}_1}(y), \ldots, \mathsf{P}_{\mathbf{Y}_{\mathsf{M}_n}}(y)\}. \tag{22}$$

It can then be shown that, if $m(y) \in [\pi - \varepsilon', \pi + \varepsilon']$, then we also have $\bar{\pi}(y) \in [\pi - \varepsilon, \pi + \varepsilon]$. We can thus upper-bound

$$\begin{aligned} &P_{[\mathsf{T}_n]}(\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)) \\ &\quad \leq |\mathcal{Y}| \max_{y \in \mathcal{Y}} P_{[\mathsf{T}_n]}\big(m(y) \notin [\pi - \varepsilon', \pi + \varepsilon']\big). \end{aligned} \tag{23}$$

We next note that, by the assumption $\mathsf{T}_n < \mathsf{M}_n/2$, the median $m(y)$ is bounded by

$$\min_{i > \mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}(y) \leq m(y) \leq \max_{i > \mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}(y). \tag{24}$$

We can thus upper-bound

$$\begin{aligned} &P_{[\mathsf{T}_n]}\big(m(y) \notin [\pi - \varepsilon', \pi + \varepsilon']\big) \\ &\quad \leq P_{[\mathsf{T}_n]}\left(\min_{i > \mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}(y) < \pi(y) - \varepsilon'\right) \\ &\qquad + P_{[\mathsf{T}_n]}\left(\max_{i > \mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}(y) > \pi(y) + \varepsilon'\right). \end{aligned} \tag{25}$$

Expressing the minimum and the maximum as unions, applying the union bound, and bounding the resulting expressions using Hoeffding's inequality, we then obtain that

$$\begin{aligned} &P_{[\mathsf{T}_n]}\big(m(y) \notin [\pi - \varepsilon', \pi + \varepsilon']\big) \\ &\quad \leq 2(\mathsf{M}_n - \mathsf{T}_n)\exp(-2n\varepsilon') \end{aligned} \tag{26}$$

which together with (23) proves the lemma.

For the full proof, see [13, App. B]. ∎

Observe that the RHS of (21) decays exponentially in n, but the error exponent vanishes as ε' tends to zero. Thus, if we reproduce the steps in the proof of Theorem 1 for the median-based estimate, then we obtain a vanishing error exponent. This contradicts the numerical results in Section V, which demonstrate that, for a finite n, the median-based test outperforms the mean-based test; cf. Fig. 3. One possible explanation for this mismatch is that the error exponent is too pessimistic, since it is dominated by the slow convergence of $\bar{\pi}$ to π. This motivates us to study the typical error exponent, defined in (14), instead.

To simplify the analysis, we characterize the typical error exponent of the two-step median-based test described at the end of Section II. Specifically, we split each sequence $\mathbf{Y}_i$ into two subsequences $\mathbf{Y}_i^{(1)}$ and $\mathbf{Y}_i^{(2)}$ that consist of the first $\lceil \rho n \rceil$ and the remaining $n - \lceil \rho n \rceil$ samples of $\mathbf{Y}_i$, respectively. The former subsequences are then used to estimate π, while the latter subsequences are used to detect the outliers. Since the observation sequences are i.i.d., the estimate $\bar{\pi}$ will then be independent of the subsequences $\mathbf{Y}_1^{(2)}, \ldots, \mathbf{Y}_{\mathsf{M}_n}^{(2)}$, which is easier to analyze. The following theorem characterizes the typical error exponent of this test.

Theorem 2: Consider the two-step median-based test described at the end of Section II, and assume that $\lim_{n\to\infty} \log(\mathsf{M}_n)/n = 0$. Further assume that the number of outlier sequences is known and satisfies $\mathsf{T}_n < \mathsf{M}_n/2$. Then, for every pair of distributions $\mu \neq \pi$, the test (12) has the typical error exponent

$$\bar{\alpha}_{\text{median}} = 2B(\pi, \mu). \tag{27}$$

Proof (Sketch): Since the typical error exponent in the universal setting cannot be larger than the error exponent in the setting where π and μ are known, it suffices to bound the typical error exponent from below.

We next write the expected value on the RHS of (14) as

$$\begin{aligned} &\mathsf{E}\left[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\right] \\ &= \mathsf{E}[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\big|\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)]P_{[\mathsf{T}_n]}\big(\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\big) \\ &\quad + \mathsf{E}[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\big|\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)]P_{[\mathsf{T}_n]}\big(\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)\big). \end{aligned} \tag{28}$$

The second expected value can be upper-bounded by zero, since the maximal error probability cannot exceed one. As for the first expected value, the exponential decay of $\epsilon(\delta_{\text{median}}|\bar{\pi})$ when $\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$ can be analyzed by following the steps in the proof of Theorem 1. It can then be shown that

$$\begin{aligned} &\underline{\lim}_{n\to\infty} -\frac{1}{n}\mathsf{E}\left[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\big|\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right] \\ &\quad \geq (1-\rho)\big(2B(\mu,\pi) - \tilde{\varepsilon}\big) \end{aligned} \tag{29}$$

(where $\underline{\lim}$ denotes the *limit inferior*) for some $\tilde{\varepsilon}$ that vanishes as ε tends to zero. Here, the factor $(1-\rho)$ is due to the fact that we detect outliers based on subsequences of length $n - \lceil \rho n \rceil$.

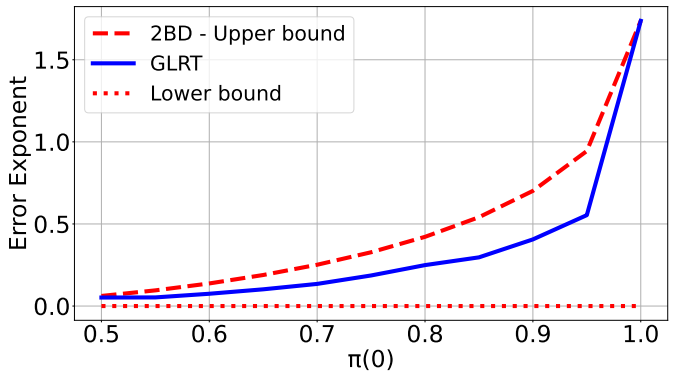

Fig. 1. Numerical demonstration that the GLRT's error exponent is generally smaller than $2B(\pi, \mu)$.

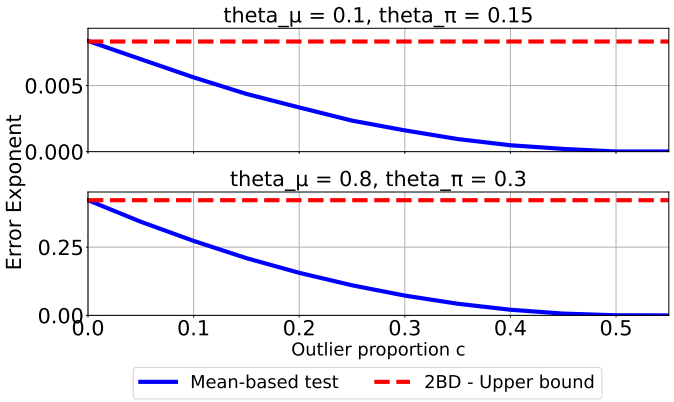

Fig. 2. Error exponent α_{mean} of the mean-based test for increasing outlier proportion c in two settings.

Lemma 1 also applies to the median-based estimate (8) evaluated for the subsequences $\mathbf{Y}_1^{(1)}, \ldots, \mathbf{Y}_{\mathsf{M}_n}^{(1)}$ by replacing n by $\lceil \rho n \rceil$. Consequently, the probability $P_{[\mathsf{T}_n]}\big(\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\big)$ tends to one as $n \to \infty$. The theorem then follows from (28) and (29) upon letting ρ and ε tend to zero.

For the full proof, see [13, App. C]. ∎

Theorem 2 only requires that M_n is subexponential in n. Thus, in contrast to Theorem 1, the theorem also applies to the case where M_n does not grow with n. As such, we can compare the typical error exponent obtained in Theorem 2 with the results obtained in [1]. Interestingly, the *typical* error exponent is equal to the exponent of the maximum-likelihood test when π and μ are known, whereas the error exponent is strictly smaller [1, Th. 9].

V. Numerical Results

We validate our theoretical findings with three experiments:

1) Numerical demonstration that the error exponent of the GLRT (4) is indeed often smaller than the typical error exponent of the two-step median-based test (Theorem 2).
2) Numerical evaluation of (20) to confirm the suboptimality of the mean-based test when $\mathsf{T}_n = c\mathsf{M}_n$.
3) Comparison of the mean-based test, the single-step median-based test,[1] and the two-step median-based test for increasing c in $\mathsf{T}_n = c\mathsf{M}_n$ and increasing M_n.

A. Suboptimality of GLRT

We select two Bernoulli distributions $\mu = B(0.7)$ and $\pi = B(\theta_\pi)$, and vary the probability $\pi(0) = 1 - \theta_\pi$ in the set $\{0.50, 0.55, \ldots, 1.00\}$. We select $\mathsf{M}_n = 4$ and numerically approximate the error exponent of the GLRT [1, Th. 2] and its lower bound [1, Th. 3]. We solve the constituting optimization problems approximately by evaluating the objective functions and constraints on a grid. More specifically, for each $\pi(0)$ we evaluate the Bernoulli distributions q_1 through q_4 in [1, eqs. (20)–(21)] for success probabilities θ_{q_i}, $i = 1, \ldots, 4$ drawn from the set $\{0.00, 0.05, \ldots, 1\}$. We proceed along similar lines to evaluate the lower bound [1, Th. 3]. The results of this simulation are depicted in Fig. 1. We observe that the GLRT's error exponent reaches $2B(\pi, \mu)$ only for $\pi(0) = 0.5$ and $\pi(0) = 1$. While the performance of the GLRT will be better for larger M_n as per [1, Th. 3], note that the lower bound from this theorem becomes vacuous for small values of M_n. Indeed, for $\mathsf{M}_n = 4$ our approximation of the lower bound evaluates to zero; cf. [1, Fig. 1].

[1]The single-step median-based test uses the same observation sequences to produce the median-based estimate $\bar{\pi}$ of π and to detect the outlier sequences.

B. Mean-based test when $\mathsf{T}_n = c\mathsf{M}_n$

With a positive outlier proportion c, the set of the observations $\mathbf{Y}_1^{\mathsf{M}_n}$ is drawn from the mixture distribution $\nu = c\mu + (1-c)\pi$. We evaluate the corresponding error exponent of the mean-based test (7) for Bernoulli distributions (i.e., $\mathcal{Y} = \{0, 1\}$) and for outlier proportions c in the set $\{0.00, 0.05, \ldots, 0.55\}$. We select the success probabilities θ_μ and θ_π of the outlier and typical distribution, respectively, from the pairs $(0.10, 0.15)$ and $(0.8, 0.3)$. Hence, we evaluate the cases where these two distributions are similar and non-similar, respectively. We evaluate (20) numerically for Bernoulli distributions Q_1 and Q_2 assuming all success probabilities in the set $\{0.001, 0.006, 0.011, \ldots, 0.996\}$. Fig. 2 shows the results for both pairs of θ_μ, θ_π. As expected, the error exponent is larger for non-similar distributions π and μ, and the error exponent of the mean-based test coincides with the upper bound of $2B(\pi, \mu)$ only if $c = 0$. (Note that in this case we assume that M_n grows superlinearly with n.) We also observe that α_{mean} decays as c increases and is zero if $c \geq 0.5$, as in this case the mean $\hat{\pi}$ becomes a better estimate of the outlier distribution μ than of the typical distribution π.

C. Non-asymptotic performance of different tests

We finally evaluate the performance of the mean-based test, and the one-step and two-step median-based tests in the non-asymptotic regime, i.e., for finite M_n and n. Specifically, we set $\mathsf{M}_n = 500$ and vary the outlier proportion c in the set $\{0.05, 0.06, \ldots, 0.5\}$. For each c, we select μ and π as categorical distributions with $|\mathcal{Y}| = 5$ and with the individual symbol probabilities drawn from a uniform distribution and

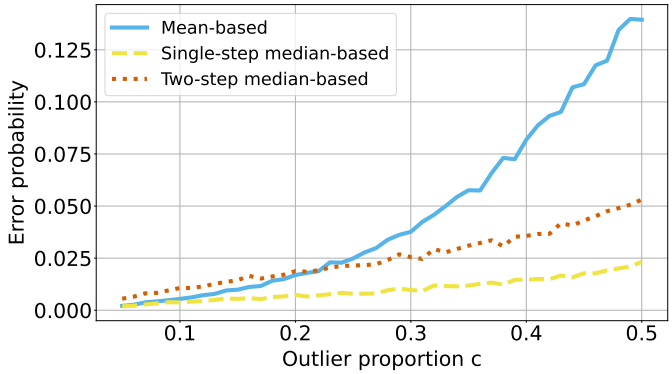

Fig. 3. Comparison of mean-based, single-step median-based, and two-step median based tests over increasing outlier proportions c.

subsequently normalized to 1. We next sample $\mathsf{T}_n = \lfloor c\mathsf{M}_n \rfloor$ and $\mathsf{M}_n - \mathsf{T}_n$ sequences of length $n = 250$ from μ and π, respectively. From these, we compute the test statistics of the mean-based, single-step median-based, and two-step median-based tests, setting $\rho = 0.5$ for the latter. Fig. 3 shows the average error probability of each test, averaged over 200 independent runs of our experiments. We observe that the mean-based test performs worst, unless c is small, which is probably due to the biased estimate of π when T_n is proportional to M_n (cf. Fig. 2). Furthermore, the single-step median-based test generally outperforms the two-step test that we analyzed in Theorem 2. We believe that this is because the two-step test suffers from poor sample efficiency. However, for the one-step median-based test, we have not been able to obtain an expression for the typical error exponent, since in this case $\bar{\pi}$ and the observation sequences $\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathsf{M}_n}$, based on which we detect outliers, are dependent.

VI. Computational Complexity

The mean-based and two-step median-based tests declare those T_n sequences as outliers for which the respective test statistics $D\left(\mathsf{P}_{\mathbf{Y}_j} \middle\| \hat{\pi}\right)$ and $D\left(\mathsf{P}_{\mathbf{Y}_j^{(2)}} \middle\| \bar{\pi}\right)$ are the largest. The tests hence require computing the type of each sequence (with a computational complexity of $\mathcal{O}(n|\mathcal{Y}|)$), taking the average or median over all types (computation complexity: $\mathcal{O}(\mathsf{M}_n|\mathcal{Y}|)$ or $\mathcal{O}(|\mathcal{Y}|\mathsf{M}_n \log \mathsf{M}_n)$, respectively), computing the test statistic for each of the M_n sequences (computational complexity: $\mathcal{O}(\mathsf{M}_n|\mathcal{Y}|)$), and sorting the M_n test statistics (computational complexity: $\mathcal{O}(\mathsf{M}_n \log \mathsf{M}_n)$) to determine the T_n largest ones. Keeping $|\mathcal{Y}|$ fixed, the optimal error exponent of $2B(\pi, \mu)$ can thus be achieved with a complexity of $\mathcal{O}(n) + \mathcal{O}(\mathsf{M}_n \log \mathsf{M}_n)$. In contrast, the GLRT proposed in [1] achieves this error exponent only asymptotically as $\mathsf{M}_n \to \infty$, but has a computational complexity of $\mathcal{O}(\mathsf{M}_n^{\mathsf{T}_n})$. Framing the test as a combinatorial clustering problem [15, Sec. 4.2], and approaching it suboptimally via K-means [16], [17], reduces the computational complexity to $\mathcal{O}(\mathsf{M}_n)$, but with an error exponent that is smaller than $2B(\pi, \mu)$ [17, Th. 1 & 3].

Acknowledgment

This work was supported in part by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI), by the Spanish Ministerio de Ciencia, Innovación y Universidades under Grant PID2024-159557OB-C21 (MICIU/AEI/10.13039/501100011033 and ERDF/UE), and by the Comunidad de Madrid under Grant IDEA-CM (TEC-2024/COM-89).

References

[1] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4066–4082, July 2014.

[2] L. Zhou, A. Hero, and Y. Wei, "Second-order asymptotically optimal outlying sequence detection with reject option," in *Proc. IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–6.

[3] S. Nitinawarat and V. V. Veeravalli, "Universal quickest outlier detection and isolation," in *Proc. 2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 770–774.

[4] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Universal outlying sequence detection for continuous observations," in *Proc. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4254–4258.

[5] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal sequential outlier hypothesis testing," in *Proc. 2014 IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 3205–3209.

[6] ——, "Universal sequential outlier hypothesis testing," in *Proc. 2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 281–285.

[7] S. C. Sreenivasan and S. Bhashyam, "Sequential nonparametric detection of anomalous data streams," *IEEE Signal Processing Letters*, vol. 28, pp. 932–936, 2021.

[8] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Sublinear algorithms for outlier detection and generalized closeness testing," in *Proc. 2014 IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 3200–3204.

[9] A. Barg and G. Forney, "Random codes: minimum distances and error exponents," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2568–2573, Sept 2002.

[10] A. Nazari, A. Anastasopoulos, and S. S. Pradhan, "Error exponent for multiple-access channels: Lower bounds," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5095–5115, Sept 2014.

[11] N. Merhav, "Error exponents of typical random codes," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6223–6235, Sept 2018.

[12] L. V. Truong and A. Guillén i Fàbregas, "Generalized random Gilbert-Varshamov codes: Typical error exponent and concentration properties," *IEEE Transactions on Information Theory*, vol. 70, no. 2, pp. 820–853, Feb 2024.

[13] B. C. Geiger, T. Koch, J. Mihaljević, and M. Toller, "Universal outlier hypothesis testing via mean- and median-based tests," 2026. [Online]. Available: https://arxiv.org/abs/2601.00712

[14] W. Hoeffding, "Probabilities inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.

[15] Y. Li and V. V. Veeravalli, "Outlying sequence detection in large datasets: Comparison of universal hypothesis testing and clustering," in *Proc. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 6180–6184.

[16] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," in *Proc. 2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 988–992.

[17] ——, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," *IEEE Transactions on Signal Processing*, vol. 67, no. 8, pp. 2115–2128, 2019.