Universal Outlier Hypothesis Testing

Yun Li, Student Member, IEEE, Sirin Nitinawarat, Member, IEEE, and Venugopal V. Veeravalli, Fellow, IEEE

Abstract

The following outlier hypothesis testing problem is studied in a universal setting. Vector observations are collected each with $M \geq 3$ coordinates, a small subset of which are outlier coordinates. When a coordinate is an outlier, the observations in that coordinate are assumed to be distributed according to an “outlier” distribution, distinct from the “typical” distribution governing the observations in all the other coordinates. Nothing is known about the outlier and 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(s). For models with exactly one outlier, a universal test based on the principle of the generalized likelihood test is proposed, and is shown to be universally exponentially consistent. A single-letter characterization of the error exponent achievable by the test is derived, and it is shown that the test is asymptotically efficient as the number of coordinates approaches infinity. When the null hypothesis with no outlier is included, a modification of this test is shown to achieve the same error exponent under each non-null hypothesis, and also consistency under the null hypothesis universally. Then, models with more than one outliers are studied in the following settings. For the setting with a known number of distinctly distributed outliers, a universally exponentially consistent test is proposed, and its achievable error exponent is characterized. The limiting error exponent achieved by the test is characterized, and the test is shown to enjoy universally asymptotically exponential consistency. For the setting with an unknown number of identically distributed outliers, a different test is shown to achieve a positive error exponent under each non-null hypothesis, and also consistency under the null hypothesis universally. When the outliers can be distinctly distributed (with their total number being unknown), it is shown that a universally exponentially consistent test cannot exist, even when the typical distribution is known and the null hypothesis is excluded.

Keywords: anomaly detection, big data, classification, detection and estimation, error exponent, fraud detection, generalized likelihood, outlier detection, outlier hypothesis testing, universal consistency, universally exponential consistency

I. INTRODUCTION

We consider the following inference problem, which we term outlier hypothesis testing. In vector observations each with a large number, say $M$, of coordinates, it is assumed that there is a small subset of outlier coordinates. Specifically, when the $i$-th coordinate is an outlier, the distribution governing the observations in that coordinate is assumed to be distinct from that governing the observations in all the “typical” coordinates. The goal is to design a test to identify all the outlier coordinates. We shall be interested in a universal setting of this problem, where the test has to perform well without any prior knowledge of the outlier and typical distributions.

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. The main differences between this outlier detection problem and our outlier hypothesis testing problem are: (i) in the former problem, the outlier observations constitute a much smaller fraction of the
entire observations than in the latter problem, and (ii) these outlier observations can be arbitrarily spread out among all observations in the outlier detection problem, whereas all the outlier observations are concentrated in a fixed subset of coordinates in the outlier hypothesis testing problem.

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 fraud and anomaly detection in large data sets \([3], [4]\), event detection and environment monitoring in sensor networks \([5]\), understanding of visual search in humans and animals \([6]\), and optimal search and target tracking \([7]\).

Universal outlier hypothesis testing is related to 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 approach is based on the principle of the generalized likelihood test \([8], [9]\). For example, in the simple-versus-composite case, the goal is to make a decision in favor of either the null distribution, which is known to the tester, or a family of alternative distributions. A fundamental result concerning the asymptotic optimality of the generalized likelihood ratio test in this case was shown in \([10]\). When some uncertainty is present in the null distribution as well, i.e., the composite-versus-composite setting, the optimality of the generalized likelihood ratio test has been examined under various conditions \([9]\).

Universal outlier hypothesis testing is also closely related to homogeneity testing and classification \([11]–[15]\). 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 to one of multiple streams of training data with distinct labels. A metric that is commonly used to quantify the performance of a test is consistency and exponential 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 \([14], [15]\), a classifier based on the principle of the generalized likelihood test was shown to be optimal under the asymptotic Neyman-Pearson criterion. In particular, in \([14]\), the classifier is designed to minimize the error probability under the homogeneous hypothesis, under a predefined constraint on the exponent for the error probability under the inhomogeneous hypothesis. And, in \([15]\), the classifier is designed to minimize the probability of rejection, under a constraint on the probability of misclassification. However, a limitation of these tests is that the aforementioned optimality is achieved only when the distribution of the test data is separated enough from those of all unmatched training data, the extent of which depends on the predefined constraint on the error exponent.

In our outlier hypothesis testing problem, we have no information regarding the outlier and typical distributions. In particular, the outliers can be arbitrarily distributed as long as each of the outlier distributions is different from the typical distribution. In addition, we have no training data to learn these distributions before the detection is performed. 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 main finding in this paper is that for our outlier hypothesis testing problem, one can construct universal tests that are far more efficient than for the other inference problems mentioned previously, such as homogeneity testing or classification. This seems quite surprising to us, as no training data is required in our test construction.

Our technical contributions are as follows. In Section III, we consider models with at most one outlier. We propose a universal test following the principle of the generalized likelihood test. When the outlier is always present and the typical distribution is known, we show that our test achieves the same optimal error exponent as if the outlier distribution is also known as well. We also show (in Example 1) that the same optimal error exponent cannot be achieved universally by any test when the typical distribution is not known. In such a completely universal setting, we prove that our test is universally exponentially consistent for each \(M \geq 3\). 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. When there is also the null hypothesis with no outlier, we show that there cannot exist a universally exponentially consistent test even when the typical distribution is known. Nevertheless, by slightly modifying our universal test, we are able to achieve the same error exponent under every hypothesis with the outlier present, and consistency under the null hypothesis. In Section IV, we consider models with multiple outliers. For models with a known number of distinctly distributed outliers, we construct a universally exponentially consistent test and show that it achieves a positive limiting error exponent as $M$ goes to infinity, whenever the limit of the optimal error exponent, when all the outlier and typical distributions are known, is positive. For models with an unknown number of identical outliers, we construct a different test that achieves a positive error exponent under every hypothesis with outliers, and consistency under the null hypothesis universally. For models with unknown number of distinct outliers, we show that there cannot exist a universally exponentially consistent test, even when the typical distribution is known and when the null hypothesis is excluded.

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 $y = y^m = (y_1, \ldots, y_m) \in \mathcal{Y}^m$, denoted by $\gamma = \gamma_y \in \mathcal{P}(\mathcal{Y})$, is defined as

$$\gamma(y) \equiv \frac{1}{m} \left| \{ k = 1, \ldots, m : y_k = y \} \right|,$$

$y \in \mathcal{Y}$.

Our results will be stated in terms of various distance metrics between a pair of distribution $p, q \in \mathcal{P}(\mathcal{Y})$. In particular, we shall consider two symmetric distance metrics: the Bhattacharyya distance and Chernoff information, denoted respectively by $B(p, q)$ and $C(p, q)$, and defined as (see, e.g., [16])

$$B(p, q) \equiv - \log \left( \sum_{y \in \mathcal{Y}} p(y)^{1/2} q(y)^{1/2} \right)$$

and

$$C(p, q) \equiv \max_{s \in [0,1]} - \log \left( \sum_{y \in \mathcal{Y}} p(y)^s q(y)^{1-s} \right),$$

respectively. Another distance metric, which will be key to our study, is the relative entropy, denoted by $D(p \parallel q)$ and defined as

$$D(p \parallel q) \equiv \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{q(y)}.$$

Unlike the Bhattacharyya distance [1] and Chernoff information [2], the relative entropy in [3] is a non-symmetric distance [16].

The following technical facts will be useful; their derivations can be found in [16] Theorem 11.1.2. 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

\[ p(y^n) = \exp \left\{ -n \left( D(\gamma \| p) + H(\gamma) \right) \right\}, \]  

(4)

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

\[ \gamma(y^n) = \exp \left\{ -nH(\gamma) \right\}. \]  

(5)

III. Universal Outlier Hypothesis Testing

A. Models with Exactly One Outlier

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 \( Y_{ik} \). 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 \( \mu \) and \( \pi \) except that \( \mu \neq \pi \), and that each of them has a 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

\[ p_i(y^{Mn}) = p_i\left( y^{(1)}, \ldots, y^{(M)} \right) = \prod_{k=1}^{n} \left\{ \mu(y^{(i)}_k) \prod_{j \neq i} \pi(y^{(j)}_k) \right\}, \]

where

\[ y^{(i)} = (y^{(i)}_1, \ldots, y^{(i)}_n), \quad i = 1, \ldots, M. \]

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 \( \delta \) is not allowed to depend on \( (\mu, \pi) \).

For a universal test, the maximal error probability, which will be a function of the test and \( (\mu, \pi) \), is

\[ e(\delta, (\mu, \pi)) \triangleq \max_{i=1,\ldots,M} \sum_{y^{Mn} \in \mathcal{Y}^{Mn}, \delta(y^{Mn}) \neq i} p_i(y^{Mn}), \]

and the corresponding error exponent is defined as

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

(6)

Throughout the paper, we consider the error exponent as \( n \) goes to infinity, while \( M \), and hence the number of hypotheses, is kept fixed. Consequently, the error exponent in (6) also coincides with the one for the average
probability of error.

A test is termed universally consistent if the maximal error probability converges to zero as the number of samples goes to infinity, i.e.,

$$e(\delta, (\mu, \pi)) \to 0,$$

for any $$(\mu, \pi)$$, $$\mu \neq \pi$$ as $$n \to \infty$$. It is termed universally exponentially consistent if the exponent for the maximal error probability is strictly positive, i.e.,

$$\alpha(\delta, (\mu, \pi)) > 0,$$

for any $$\mu \neq \pi$$.

1) Proposed Universal Test: We now describe our universal test in two setups when only $$\pi$$ is known, and when neither $$\mu$$ nor $$\pi$$ is known, respectively. Our test follows the principle of the generalized likelihood test [8], [9].

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

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

$$= - \left[ H(\gamma_i) + D(\gamma_i \| \mu) \right] - (M-1) \left[ H \left( \frac{\sum_{j \neq i} \gamma_j}{M-1} \right) + D \left( \frac{\sum_{j \neq i} \gamma_j}{M-1} \| \pi \right) \right]$$

$$= - \left[ H(\gamma_i) + D(\gamma_i \| \mu) \right] - \sum_{j \neq i} \left[ H(\gamma_j) + D(\gamma_j \| \pi) \right],$$

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

When $$\pi$$ is known and $$\mu$$ is unknown, we compute the generalized log-likelihood of $$y^M$$ by replacing $$\mu$$ in (10) with its maximum likelihood (ML) estimate $$\hat{\mu}_i \triangleq \gamma_i$$, $$i = 1, \ldots, M$$, as

$$L^{\text{typ}}_i(y^M) = -H(\gamma_i) - \sum_{j \neq i} \left[ H(\gamma_j) + D(\gamma_j \| \pi) \right].$$

Similarly, when neither $$\mu$$ nor $$\pi$$ is known, we compute the generalized log-likelihood of $$y^M$$ by replacing the $$\mu$$ and $$\pi$$ in (9) and (10) with their maximum likelihood (ML) estimates $$\hat{\mu}_i \triangleq \gamma_i$$, and $$\hat{\pi}_i \triangleq \frac{\sum_{j \neq i} \gamma_j}{M-1}$$, $$i = 1, \ldots, M$$, as

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

Finally, we decide upon the coordinate with the largest generalized log-likelihood to be the outlier. Using (11), (12), our universal tests in the two cases can be described respectively as

$$\delta(y^M) = \arg\min_{i=1, \ldots, M} L^{\text{typ}}_i(y^M),$$

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

$$U^{\text{typ}}_i(y^M) \triangleq \sum_{j \neq i} D(\gamma_j \| \pi),$$

and

$$\delta(y^M) = \arg\min_{i=1, \ldots, M} L^{\text{univ}}_i(y^M),$$

(13)

(14)

(15)
when neither \( \mu \) nor \( \pi \) is known, where for each \( i = 1, \ldots, M \),
\[
U_i^{\text{univ}}(y^{Mn}) \triangleq \sum_{j \neq i} D \left( \gamma_j \| \frac{\sum_{k \neq i} \gamma_k}{M-1} \right).
\]  
(16)

In (13) and (15), should there be multiple minimizers, we pick one of them arbitrarily.

2) \textbf{Results:} Our first theorem for models with one outlier characterizes the optimal exponent for the maximal error probability when both \( \mu \) and \( \pi \) are known, and when only \( \pi \) is known.

\textbf{Theorem 1.} When \( \mu \) and \( \pi \) are both known, the optimal exponent for the maximal error probability is equal to
\[
2B(\mu, \pi).
\]  
(17)

Furthermore, the error exponent in (17) is achievable by a test that uses only the knowledge of \( \pi \). In particular, such a test is our proposed test in (13), (14).

\textbf{Remark 1.} It is interesting to note that when only \( \mu \) is known, one can also achieve the optimal error exponent in (17). However, we do not yet know if the corresponding version of our proposed test, wherein the \( \pi \) in (10) is replaced with \( \hat{\pi}_i = \frac{\sum_{k \neq i} \gamma_k}{M-1} \), \( i = 1, \ldots, M \), is optimal. Nevertheless, a different test will be presented in Appendix A and will be shown to achieve the optimal error exponent in (17).

Consequently, in the completely universal setting, when nothing is known about \( \mu \) and \( \pi \) except that \( \mu \neq \pi \), and both \( \mu \) and \( \pi \) have full supports, it holds that for any universal test \( \delta \),
\[
\alpha(\delta, (\mu, \pi)) \leq 2B(\mu, \pi).
\]  
(18)

Given the second assertion in Theorem 1 it might be tempting to think that it would be possible to design a test to achieve the optimal error exponent of \( 2B(\mu, \pi) \) universally when neither \( \mu \) nor \( \pi \) is known. Our first example shows that such a goal cannot be fulfilled, and hence we need to be content with a lesser goal.

\textbf{Example 1:} Consider the model with \( M = 3 \), and a distinct pair of distributions \( p \neq \overline{p} \) on \( \mathcal{Y} \) with full supports. We now show that there cannot exist a universal test that achieves the optimal error exponent of \( 2B(\mu, \pi) \) even just for the two models when \( \mu = p, \pi = \overline{p} \) and when \( \mu = \overline{p}, \pi = p \), both of which have \( 2B(\mu, \pi) = 2B(p, \overline{p}) \).

To this end, let us look at the region when a universal test \( \delta \) decides that the first coordinate is the outlier, i.e.,
\[
A_1 = \{ y^{3n} : \delta (y^{3n}) = 1 \}. \]

Let \( \mathbb{P}_{p, \overline{p}, \overline{p}} \) denote the distribution corresponding to the first hypothesis of the first model, i.e., when \( y^{(1)} \) are i.i.d. according to \( p \), and \( y^{(2)} \) and \( y^{(3)} \) are i.i.d. according to \( \overline{p} \). Similarly, let \( \mathbb{P}_{p, \overline{p}, p} \) denote the distribution corresponding to the second hypothesis of the second model, i.e., when \( y^{(2)} \) are i.i.d. according to \( \overline{p} \), and \( y^{(1)} \) and \( y^{(3)} \) are i.i.d. according to \( p \). Suppose that \( \delta \) achieves the best error exponent of \( 2B(p, \overline{p}) \) for the first model when \( \mu = p, \pi = \overline{p} \). Then, it must holds that
\[
\lim_{n \to \infty} - \frac{1}{n} \log \mathbb{P}_{p, \overline{p}, \overline{p}} \{ A_1 \} \geq 2B(p, \overline{p}).
\]  
(19)

It now follows from the classic result of Hoeffding [10] in binary hypothesis testing that (19) implies that
\[
\lim_{n \to \infty} - \frac{1}{n} \log \mathbb{P}_{p, \overline{p}, p} \{ A_1 \} \leq \left( \min_{D(q(y_1) \| p) + D(q(y_2) \| \overline{p}) + D(q(y_3) \| p) \leq 2B(p, \overline{p})} D(q(y_1) \| p) + D(q(y_2) \| \overline{p}) + D(q(y_3) \| p) \right)^+ \leq \left( \min_{D(q(y_1) \| p) + D(q(y_2) \| \overline{p}) + D(q(y_3) \| p) \leq 2B(p, \overline{p})} 2B(p, \overline{p}) + D(q(y_2) \| p) - D(q(y_3) \| \overline{p}) \right)^+ \leq (2B(p, \overline{p}) - D(p \| \overline{p}))^+ = 0,
\]  
(20)
where the last equality follows from our Lemma 2 below. Consequently, the test cannot yield even a positive error exponent for the second model when \( \mu = \overline{\mu}, \pi = p \).

Example 1 shows explicitly that when neither \( \mu \) nor \( \pi \) is known, it is impossible to construct a test that achieves \( 2B(\mu, \pi) \) universally. In fact, the example shows that had we insisted on achieving the best error exponent of \( 2B(\mu, \pi) \) for some pairs of \( \mu, \pi \), it might not be possible to achieve even positive error exponents for some other pairs of \( \mu, \pi \). This motivates us to seek instead a test that yields just a positive (no matter how small) error exponent \( \alpha(\delta, (\mu, \pi)) > 0 \) for every \( \mu, \pi, \mu \neq \pi \), i.e., achieving universally exponential consistency. We argue that without knowing either \( \mu \) or \( \pi \), it is not even clear that such a lesser objective could be fulfilled. It might be true that \( \frac{1}{M}\mu + \frac{M-1}{M}\pi \) can be efficiently estimated by just averaging the observations in all coordinates. However, since \( \mu \) is unknown and can be any distribution (distinct from \( \pi \)), there is no universal way to use that estimate to efficiently guess \( \pi \). One of our main contributions in this paper is to show that our proposed universal test in (15), (16) is universally exponentially consistent. Furthermore, for every pair of distributions \( \mu, \pi, \mu \neq \pi \), it holds that

\[
\alpha(\delta, (\mu, \pi)) = \min_{q_1, \ldots, q_M} D(q_1 || \mu) + D(q_2 || \pi) + \ldots + D(q_M || \pi),
\]

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

\[
\sum_{j \neq 1} D(q_j \| \frac{1}{M} \sum_{k=1}^{M-1} q_k) \geq \sum_{j \neq 2} D(q_j \| \frac{1}{M} \sum_{k=1}^{M-1} q_k).
\]

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

\[
\lim_{n \to \infty} \mathbb{P}_t \left\{ \left\| \frac{1}{M} \sum_{j=1}^{M} \gamma_j - \left( \frac{1}{M}\mu + \frac{M-1}{M}\pi \right) \right\|_1 > \epsilon \right\} = 0,
\]

where \( \| \cdot \|_1 \) denotes the 1-norm of the argument distribution. Since \( \frac{1}{M}\mu + \frac{M-1}{M}\pi \to \pi \) as \( M \to \infty \), 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 (achievable error exponent) of our proposed universal test in (15), (16) when \( M \to \infty \) (after having taken the limit as \( n \) goes to infinity).

Our last result for models with one outlier shows that in the completely universal setting, as \( M \to \infty \), our proposed universal test in (15), (16) achieves the optimal error exponent in (17) corresponding to the case in which both \( \mu \) and \( \pi \) are known.

**Theorem 3.** For each \( M \geq 3 \), the exponent for the maximal error probability achievable by our proposed universal test \( \delta \) in (15), (16) is lower bounded by

\[
\min_{q \in \mathcal{P}(\mathcal{Y})} D(q || \pi) \leq \frac{1}{\pi^*} \left( 2B(\mu, \pi) + C^* \right)
\]

where \( C^* \equiv -\log \left( \min_{y \in \mathcal{Y}} \pi(y) \right) \leq \infty \) by the fact that \( \pi \) has a full support.

The lower bound for the error exponent in (24) is nondecreasing in \( M \geq 3 \). Furthermore, as \( M \to \infty \), this lower bound converges to the optimal error exponent \( 2B(\mu, \pi) \); hence, our test is asymptotically efficient by which
it means that
\[
\lim_{M \to \infty} \alpha(\delta, (\mu, \pi)) = 2B(\mu, \pi),
\]
which from Theorem 1 is equal to the optimal error exponent when both \(\mu\) and \(\pi\) are known.

**Example 2:** We now provide some numerical results for an example with \(\mathcal{Y} = \{0, 1\}\). Specifically, the three plots in the figure below are for three pairs outlier and typical distributions being \(\mu = (p(0) = 0.3, p(1) = 0.7), \pi = (0.7, 0.3); \mu = (0.35, 0.65), \pi = (0.65, 0.35);\) and \(\mu = (0.4, 0.6), \pi = (0.6, 0.4),\) respectively. Each horizontal line corresponds to \(2B(\mu, \pi)\), and each curve line corresponds to the lower bound in (24) for the error exponent achievable by our proposed universal test. As shown in these plots, the lower bounds converge to \(2B(\mu, \pi)\) as \(M \to \infty\), i.e., our proposed universal test is asymptotically optimal for all three pairs \(\mu, \pi\).

![Graph](image.png)

**B. Models with At Most One Outlier**

A natural question that arises at this point is what would happen if it is also possible that no outlier is present. To answer this question, we now consider models that append an additional null hypothesis with no outlier to the previous models consider in Section III-A. In particular, under the null hypothesis, the joint distribution of all the observations is given by
\[
p_0(y^{Mn}) = \prod_{k=1}^{n} \prod_{i=1}^{M} \pi(y^{(i)}).\]

A universal test \(\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}\) will now also accommodate for an additional decision for the null hypothesis. Correspondingly, the maximal error probability is now computed with the inclusion of the null hypothesis according to
\[
e(\delta, (\mu, \pi)) \triangleq \max_{i=0,1,\ldots,M} \sum_{y^{Mn} : \delta(y^{Mn}) \neq i} p_i(y^{Mn}).\]

With just one additional null hypothesis, contrary to the previous models with one outlier, it becomes impossible to achieve universally exponential consistency even with the knowledge of the typical distribution. This pessimistic
result reaffirms that our previous finding that universally exponential consistency is attained for the models with one outlier is indeed quite surprising.

**Proposition 4.** For the setting with the additional null hypothesis, there cannot exist a universally exponentially consistent test even when the typical distribution is known.

In typical applications such as environment monitoring and fraud detection, the consequence of a missed detection of the outlier can be much more catastrophic than that of a false positive. In addition, Proposition 4 tells us that there cannot exist a universal test that yields exponential decays for both the conditional probability of false positive (under the null hypothesis) and the conditional probabilities of missed detection (under all non-null hypotheses). Consequently, it is natural to look for a universal test fulfilling a lesser objective: attaining universally exponential consistency for conditional error probabilities under only all the non-null hypotheses, while seeking only universal consistency for the conditional error probability under the null hypothesis. We now show that such a test can be obtained by slightly modifying our earlier test. Furthermore, in addition to achieving universal consistency under the null hypothesis, this new test achieves the same exponent as in (21), (22) in Theorem 2 for the conditional error probabilities under all non-null hypotheses.

In particular, we modify our previous test in (15), (16) to allow for the possibility of deciding for the null hypothesis as follows.

\[
\delta(y_{MN}) = \begin{cases} 
\arg \min_{i=1,\ldots,M} U_i(y_{MN}), & \text{if } \max_{j\neq k} (U_j(y_{MN}) - U_k(y_{MN})) > \lambda_n, \\
0, & \text{otherwise},
\end{cases}
\]

where \(\lambda_n = O(\log \frac{n}{n})\) and the ties in the first case of (26) are broken arbitrarily.

**Theorem 5.** For every pair of distributions \(\mu, \pi, \mu \neq \pi\), our test in (26) yields a positive exponent for the conditional probability of error under every non-null hypothesis \(i = 1, \ldots, M\), and a vanishing conditional probability of error under the null hypothesis. In particular, the achievable error exponent under every non-null hypothesis is the same as that given in (21), (22), i.e., for each \(i = 1, \ldots, M\), our test in (26) achieves

\[
\lim_{n \to \infty} -\frac{1}{n} \log \left( \mathbb{P}_i \{ \delta \neq i \} \right) = \min_{q_1,\ldots,q_M} D(q_1\|\mu) + D(q_2\|\pi) + \ldots + D(q_M\|\pi),
\]

where the minimum above is over the set of \((q_1,\ldots,q_M)\) satisfying (22). In addition, it also yields that

\[
\lim_{n \to \infty} \mathbb{P}_0 \{ \delta \neq 0 \} = 0.
\]

Since under every non-null hypothesis, our modified test in (26) achieves the same exponent (the value of the optimization problem in (21), (22)) for the conditional error probability as our previous test in (15), (16) when the null hypothesis is absent, we get the following corollary by just observing that Theorem 3 was proved by finding a suitable lower bound for the value of the optimization problem in (21), (22).

**Corollary 6.** For each \(M \geq 3\) and under every non-null hypothesis \(i = 1, \ldots, M\), the exponent for the conditional error probability achievable by our test in (26) is lower bounded as

\[
\lim_{n \to \infty} -\frac{1}{n} \log \left( \mathbb{P}_i \{ \delta \neq i \} \right) \geq \min_{q \in \mathcal{P}(Y)} 2B(\mu, q),
\]

where \(D(q\|\pi) \leq \frac{1}{2}(2B(\mu,\pi) + C_n)\).
where $C_{\pi} \triangleq -\log \left( \min_{y \in \mathcal{Y}} \pi(y) \right) < \infty$. Consequently, as $M \to \infty$, this lower bound converges to the optimal error exponent $2B(\mu, \pi)$, i.e., for every $i = 1, \ldots, M$, our test in (29) achieves
\[
\lim_{M \to \infty} \lim_{n \to \infty} \frac{1}{n} \log \left( \mathbb{P}_1 \{ \delta \neq i \} \right) = 2B(\mu, \pi),
\]
while also yielding that
\[
\lim_{n \to \infty} \mathbb{P}_0 \{ \delta \neq 0 \} = 0.
\]

IV. UNIVERSAL MULTIPLE OUTLIER HYPOTHESIS TESTING

We now generalize our results in the previous section to models with multiple outliers. With more than one outlier, it may be more natural to consider models for which the different outlier coordinates are distinctly distributed, and therefore our models will allow for this possibility. Also, we shall consider two settings: the setting with a known number, say $T > 1$, of outliers, and the setting with an unknown number of outliers, say up to $T$ outliers.

We start by describing a generic model with possibly distinctly distributed outliers, the number of which is not known. As in the previous section, we denote the $k$-th observation of the $i$-th coordinate by $Y_k^{(i)} \in \mathcal{Y}, i = 1, \ldots, M, k = 1, \ldots, n$. Most of the coordinates are commonly distributed according to the “typical” distribution $\pi \in \mathcal{P}(\mathcal{Y})$ except for a small (possibly empty) subset $S \subset \{1, \ldots, M\}$ of “outlier” coordinates, each of which is assumed to be distributed according to an outlier distribution $\mu_i, i \in S$. Nothing is known about $\{\mu_i\}_{i=1}^M$ and $\pi$ except that each $\mu_i \neq \pi, i = 1, \ldots, M$, and that all $\mu_i, i = 1, \ldots, M$, and $\pi$ have full supports. In the following presentation, we sometimes consider the special case when all the outlier coordinates are identically distributed, i.e., $\mu_i = \mu, i = 1, \ldots, M$.

For a hypothesis corresponding to an outlier subset $S \subset \{1, \ldots, M\}$, $|S| < \frac{M}{2}$, the joint distribution of all the observations is given by
\[
\begin{align*}
p_S \left( y_{Mn}^{y_{Mn}} \right) &= p_S \left( y^{(1)}, \ldots, y^{(M)} \right) \\
&= \prod_{k=1}^{n} \left\{ \prod_{i \in S} \mu_i \left( y_k^{(i)} \right) \prod_{j \notin S} \pi \left( y_k^{(j)} \right) \right\},
\end{align*}
\]
where
\[
y^{(i)} = \left( y_1^{(i)}, \ldots, y_n^{(i)} \right), i = 1, \ldots, M.
\]

We refer to the unique hypothesis corresponding to the case with no outlier, i.e., $S = \emptyset$, as the null hypothesis. In the following subsections, we shall consider different settings, each being described by a suitable set $S$ comprising all possible outlier subsets.

The test for the outlier subset is done based on a universal rule $\delta : \mathcal{Y}^{Mn} \to S$. In particular, the test $\delta$ is not allowed to depend on $\left( \{\mu_i\}_{i=1}^M, \pi \right)$.

For a universal test, the maximal error probability, which will be a function of the test and $\left( \{\mu_i\}_{i=1}^M, \pi \right)$, is
\[
e \left( \delta, \left( \{\mu_i\}_{i=1}^M, \pi \right) \right) \triangleq \max_{S \in \mathcal{S}} \left\{ \sum_{y^{Mn} : \delta(y^{Mn}) \neq S} p_S(y^{Mn}) \right\},
\]
and the corresponding error exponent is defined as
\[
\alpha \left( \delta, \left( \{\mu_i\}_{i=1}^M, \pi \right) \right) \triangleq \lim_{n \to \infty} \frac{1}{n} \log e \left( \delta, \left( \{\mu_i\}_{i=1}^M, \pi \right) \right).
\]
A universal test $\delta$ is termed *universally exponentially consistent* if for every $\mu_i$, $i = 1, \ldots, M$, $\mu_i \neq \pi$, it holds that
\[
\alpha \left( \delta, \left( \{\mu_i\}_{i=1}^M, \pi \right) \right) > 0.
\]

### A. Models with Known Number of Outliers

We start by considering the case in which the number of outliers, denoted by $T > 1$, is known at the outset, i.e., $|S| = T$, for every $S \in \mathcal{S}$. Unlike the model in Section III where the outlier coordinate is always distributed according to a fixed distribution $\mu \neq \pi$ regardless of its identity $i = 1, \ldots, M$, in our model for this subsection, the distributions of different outliers $\mu_i$, $i \in S$, can vary across their coordinate indices, $i \in S$.

1) Proposed Universal Test: Note that the universal test proposed in Section III-A1 for the single outlier case, under the setup when only $\pi$ is known as well as the completely universal setup, depends only on the empirical distributions of all the $M$ coordinates (and on $\pi$ in the former setup), but not on the outlier distribution. Consequently, we can directly generalize this test to our current models (with distinct and more than one outlier). We now give a summary of this test for the current models with multiple outliers under both the setup when only $\pi$ is known and the completely universal setup.

The test is done based on the following statistics:

\[
U^{\text{typ}}_S (y^{Mn}) \triangleq \sum_{j \notin S} D(\gamma_j \| \pi),
\]

and

\[
U^{\text{univ}}_S (y^{Mn}) \triangleq \sum_{j \notin S} D \left( \gamma_j \left\| \frac{\sum_{k \in S} \gamma_k}{M - T} \right\| \right),
\]

respectively. The test then selects the hypothesis with the smallest value of such statistic (ties are broken arbitrarily), i.e.,

\[
\delta(y^{Mn}) = \arg\min_{S \subseteq \{1, \ldots, M\}, \ |S| = T} U^{\text{typ}}_S (y^{Mn})
\]

for the setting when only $\pi$ is known, and

\[
\delta(y^{Mn}) = \arg\min_{S \subseteq \{1, \ldots, M\}, \ |S| = T} U^{\text{univ}}_S (y^{Mn}),
\]

for the completely universal setting, respectively.

2) Results:

**Proposition 7.** For every fixed number of outliers $1 < T < \frac{M}{2}$, when all the $\mu_i$, $i = 1, \ldots, M$, and $\pi$ are known, the optimal error exponent is equal to

\[
\min_{1 \leq i < j \leq M} C (\mu_i (y) \pi (y'), \pi (y) \mu_j (y')).
\]

When all outlier coordinates are identically distributed, i.e., $\mu_i = \mu \neq \pi$, $i = 1, \ldots, M$, this optimal error exponent is independent of $M$ and is equal to

\[
2B (\mu, \pi).
\]
Theorem 8. For every fixed number of outliers $1 < T < \frac{M}{2}$, when only $\pi$ is known but none of $\mu_i$, $i = 1, \ldots, M$ is known, the error exponent achievable by our proposed test in (31), (33) is equal to
\[
\min_{1 \leq i \leq M} 2B(\mu_i, \pi). \tag{37}
\]
When all outlier coordinates are identically distributed, i.e., $\mu_i = \mu, i = 1, \ldots, M$, this achievable error exponent is equal to
\[
2B(\mu, \pi), \tag{38}
\]
which, from Proposition 7, is the optimal error exponent when $\mu$ is also known.

Remark 2. Since the tester in Proposition 7 is more capable than that in Theorem 8, the optimal error exponent in (35) must be no smaller than that in (37). This is verified simply by noting that for every $i, j, 1 \leq i < j \leq M$, we get from (2) that
\[
C(\mu_i(y) \pi(y'), \pi(y) \mu_j(y')) = \max_{s \in [0, 1]} -\log \left( \sum_{y, y' \in \mathcal{Y} \times \mathcal{Y}} (\mu_i(y) \pi(y'))^s (\pi(y) \mu_j(y'))^{1-s} \right)
\geq B(\mu_i, \pi) + B(\mu_j, \pi)
\geq \min (2B(\mu_i, \pi), 2B(\mu_j, \pi)). \tag{39}
\]

As in Section III for the current models, a test $\delta$ is universally exponentially consistent if for every $\mu_i, i = 1, \ldots, M, \mu_i \neq \pi$, it holds that $\alpha(\delta, \{\mu_i\}_{i=1}^M, \pi) > 0$.

Theorem 9. For every fixed number of outliers $1 < T < \frac{M}{2}$, our proposed test $\delta$ in (32), (34) is universally exponentially consistent. Furthermore, for every $\{\mu_i\}_{i=1}^M, \mu_i \neq \pi, i = 1, \ldots, M$, it holds that
\[
\alpha(\delta, \{\mu_i\}_{i=1}^M, \pi) = \min_{S, S' \subset \{1, \ldots, M\}, |S| = T} \min_{q_1, \ldots, q_M} \sum_{i \in S} D(q_i \| \mu_i) + \sum_{j \notin S} D(q_j \| \mu_j), \tag{40}
\]
where the inner minimum above is over the set of $(q_1, \ldots, q_M)$ such that
\[
\sum_{i \notin S} D(q_i \| \frac{\sum_{k \in S} q_k}{M-T}) \geq \sum_{i \notin S'} D(q_i \| \frac{\sum_{k \in S} q_k}{M-T}). \tag{41}
\]

Note that universally exponential consistency does not imply that
\[
\lim_{M \to \infty} \alpha(\delta, \{\mu_i\}_{i=1}^M, \pi) > 0. \tag{42}
\]
Furthermore, it follows from Proposition 7 that (42) is not possible for $\{\mu_i\}_{i=1}^M, \pi$ such that
\[
\lim_{M \to \infty} \min_{1 \leq i \leq j \leq M} C(\mu_i(y) \pi(y'), \pi(y) \mu_j(y')) = 0. \tag{43}
\]
A test that satisfies (42) whenever (43) does not hold is said to enjoy universally asymptotically exponential consistency.
Theorem 10. For every $M \geq 3$, and every fixed number of outliers $1 < T < \frac{M}{2}$, the error exponent achievable by our proposed test in (32), (34) is lower bounded by

$$\min_{q \in P(Y)} D(q\|p) \leq \frac{1}{M-T} \left( \min_{i \neq j} C_i \right)$$

where $C_i \triangleq -\log \left( \min_{y \in Y} \pi(y) \right) < \infty$.

Furthermore, the proposed test enjoys universally asymptotically exponential consistency. In particular, as $M \to \infty$, the error exponent achievable by our test in (32), (34) converges as

$$\lim_{M \to \infty} \alpha \left( \delta, \left( \{\mu_i\}_{i=1}^M, \pi \right) \right) = \lim_{M \to \infty} \min_{i=1,\ldots,M} 2B(\mu_i, \pi),$$

which from (37) of Theorem 8 is also the limit of the achievable error exponent of the test in (31), (33) using the knowledge of the typical distribution.

When all outlier coordinates are identically distributed, i.e., $\mu_i = \mu \neq \pi$, $i = 1, \ldots, M$, our test is asymptotically efficient, i.e.,

$$\lim_{M \to \infty} \alpha \left( \delta, (\mu, \pi) \right) = 2B(\mu, \pi).$$

B. Models with Unknown Number of Outliers

In this section, we look at the case in which not all hypotheses in $S$ have the same number of outliers, i.e., there is uncertainty in the number of outliers in the hypothesis testing problem. Throughout this section, for a fixed number of outliers $k = 0, 1, 2, \ldots$, we assume that $S$ either contains all hypotheses with $k$ outliers, or none of them.

1) Models with Identical Outliers: We now show that when all outlier coordinates are identically distributed, even without knowing the number of outliers exactly (assumed in Section IV-A), we can still construct a universally exponentially consistent test if we know that there are always some outliers.

To present a derivation of such a test, we shall follow the principle of the generalized likelihood test similar to that in (32), (34) but now with the assumption of identical outliers being taken strictly. Then it can be shown that the negative of the generalized log-likelihood of $y^M$ for an outlier subset $S \in S$ with identical outliers, denoted by $\tilde{U}_{S}^{\text{univ}}(y^M)$, is equivalent to

$$\tilde{U}_{S}^{\text{univ}}(y^M) \triangleq \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S} \gamma_k}{T} \right\| \right) + \sum_{j \notin S} D \left( \gamma_j \left\| \frac{\sum_{k \notin S} \gamma_k}{M-T} \right\| \right),$$

and, hence, our universal test can be described (with arbitrarily broken ties) as

$$\delta(y^M) = \arg\min_{S \in S} \tilde{U}_{S}^{\text{univ}}(y^M).$$

Theorem 11. When all the outlier coordinates are identically distributed, our proposed test in (47), (48) is universally exponentially consistent for every hypothesis set excluding the null hypothesis. On the other hand, when the hypothesis set contains the null hypothesis, there cannot exist a universally exponentially consistent test even when the typical distribution is known.

Remark 3. When the null hypothesis is present, we can make a suitable modification to our test (47), (48) similar to (26) to get a universal test that achieves a positive exponent for every conditional error probability, conditioned on any non-null hypothesis, and additional consistency under the null hypothesis.
Remark 4. It is interesting to note the difference between the two statistics in (47) and (32). In particular, the statistic in (32) is equivalent to the negative of the generalized log-likelihood with distinct outliers, whereas (47) is equivalent to the negative of the generalized log-likelihood with identical outliers. For models with identical outliers, although (32) does not follow the principle of the generalized log-likelihood test as strictly as (47), we did prove that when the number of identical outliers is known, the test based on (32) is universally exponentially consistent (in Theorem 9) and is asymptotically efficient (in (46)). We suspect that a universal test based on (47) should also be asymptotically efficient for models with a known number of identical outliers. Surprisingly, the proof of this assertion seems to be more difficult than that for the test based on (32), and still eludes us. This conundrum arises partly because of the more complicated form in (47) than in (32). Nevertheless, a major drawback of the test based on (32) is that it is not universally exponentially consistent for the current models in Theorem 11 when the number of outliers is not known.

2) Models with Distinct Outliers: Our last pessimistic result shows that when the outlier coordinates can be distinctly distributed in an arbitrary manner, the assumption of a known number of outliers adopted in Section IV-A is indeed critical, as a lack thereof would make it impossible to construct a universally exponentially consistent test even when there are always some outlier coordinates.

Theorem 12. When the outlier coordinates can be distinctly distributed, there cannot exist a universally exponentially consistent test, even when the typical distribution is known and when the null hypothesis is excluded, i.e., there are always some outlier coordinates (regardless of the hypothesis).

Remark 5. The negative result in Theorem 12 should not be taken with extreme pessimism. It should be viewed as a theoretical result which holds only when each of the outliers can be arbitrarily distributed, as long as they differ from the typical distribution. In practice, there will likely be modeling constraints that would confine the set of all possible tuples of the distributions of all outliers. An extreme case of such constraints is when all the outliers are forced to be identically distributed, which is when universally exponential consistency is indeed attained (cf. Theorem 11) if the null hypothesis is excluded. An interesting future research direction would be to characterize the “least” stringent joint constraint on the distributions of the outliers that still allows us to construct universally exponentially consistent tests and is more appealing to a broader class of applications.

V. CONCLUSION

In this paper, we formulated and studied the problem of outlier hypothesis testing in a completely universal setting. Our main contribution was in constructing universal tests for this problem, and proving that these tests yield exponentially decaying probability of error under various settings. In particular, for the case with at most one outlier, we proposed a test that is universally exponentially consistent if the null hypothesis with no outlier is excluded. We also provided a characterization of the error exponent achievable by our test for each \( M \geq 3 \). Surprisingly our test is not only universally exponentially consistent, but also asymptotically efficient as the number of coordinates goes to infinity. Specifically, as \( M \) goes to infinity, the error exponent achievable by our universal test converges to the absolutely optimal error exponent when both the outlier and typical distributions are known. When there is an additional null hypothesis, a suitable modification of our test was shown to achieve exponential consistency under each hypothesis with the outlier, and consistency under the null hypothesis universally. Under every non-null hypothesis, this modified test achieves the same error exponent as that achievable when the null
hypothesis is excluded. We then extended our models to cover the case with more than one outliers. For models with a known number of outliers, the distributions of the outliers could be distinct as long as each of them differs from the typical distribution. A test was constructed and was shown to be universally exponentially consistent. Furthermore, we characterized the limiting error exponent achieved by our test, and established its universally asymptotically exponential consistency. When the number of outliers is not known, it was shown that the assumption of the outliers being identically distributed and the exclusion of the null hypothesis were both essential for existence of universally exponentially consistent test. In particular, for models with an unknown number of \textit{identically} distributed outliers, we proposed a test that is universally exponentially consistent when the null hypothesis is excluded. When the null hypothesis is included, a different test was shown to achieve a positive error exponent under every non-null hypothesis, and also consistency under the null hypothesis universally. For models with an unknown number of \textit{distinctly} distributed outliers, it was shown that even when the typical distribution is known for inter-

VI. Acknowledgments

The authors thank Dr. Jean-Francois Chamberland-Tremblay, Aly El Gamal and Dr. Maxim Raginsky for interesting discussions related to this paper.

APPENDIX

Our proofs rely on the following lemmas.

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

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

\textbf{Proof:} Let $E$ be the set of all joint distributions in $\mathcal{P}(\mathcal{Y}^J)$ with the tuple of their corresponding marginal distributions lying in $E$. It now follows from the closeness of $E$ in $\mathcal{P}(\mathcal{Y}^J)$ and the compactness of $\mathcal{P}(\mathcal{Y}^J)$ that $E$ is also closed in $\mathcal{P}(\mathcal{Y}^J)$. Let $A_n$ be the set of all $J$ tuples $(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}) = \left( \left( y_{(1)}^{(1)}, \ldots, y_{(1)}^{(J)} \right), \ldots, \left( y_{(n)}^{(1)}, \ldots, y_{(n)}^{(J)} \right) \right) \in \mathcal{Y}^{Jn}$ whose empirical distribution lies in a closed set $E \in \mathcal{P}(\mathcal{Y}^J)$. The lemma then follows by observing that $\mathbb{P} \left\{ \left( \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(J)} \right) \in A_n \right\} = \mathbb{P} \left\{ \left( \left( y_{(1)}^{(1)}, \ldots, y_{(1)}^{(J)} \right), \ldots, \left( y_{(n)}^{(1)}, \ldots, y_{(n)}^{(J)} \right) \right) \in A_n \right\}$, and by invoking the Sanov’s lemma to compute the exponent of the latter probability, i.e.,

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P} \left\{ \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(J)} \in A_n \right\} = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P} \left\{ \left( \left( y_{(1)}^{(1)}, \ldots, y_{(1)}^{(J)} \right), \ldots, \left( y_{(n)}^{(1)}, \ldots, y_{(n)}^{(J)} \right) \right) \in A_n \right\}$$

$$= \min_{q \in E} D(q \| p_1 \times \ldots \times p_J)$$

$$= \min_{(q_1, \ldots, q_J) \in E} \sum_{j=1}^{J} D(q_j \| p_j)$$

\hfill \blacksquare

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

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

(50)
In particular, the minimum on the right side of (50) is achieved by
\[
q^* = \frac{p_1^\dagger(y)p_2^\dagger(y)}{\sum_{y \in \mathcal{Y}} p_1^\dagger(y)p_2^\dagger(y)}, \quad y \in \mathcal{Y}.
\] (51)

**Proof:** It follows from the concavity of the logarithm function that
\[
D(q\parallel p_1) + D(q\parallel p_2) = \sum_{y \in \mathcal{Y}} q(y) \log \frac{q^2(y)}{p_1(y)p_2(y)}
\]
\[
= -2 \sum_{y \in \mathcal{Y}} q(y) \log \frac{p_1^\dagger(y)p_2^\dagger(y)}{q(y)}
\]
\[
\geq -2 \log \left( \sum_{y \in \mathcal{Y}} p_1^\dagger(y)p_2^\dagger(y) \right)
\] (52)

In particular, equality is achieved in (52) by \(q(y) = q^*(y)\) in (51).

It is interesting to note that from (50), we recover the known inequality discovered in [17]:
\[
2B(p_1, p_2) \leq \min \left( D(p_2\parallel p_1), D(p_1\parallel p_2) \right),
\] (53)
by evaluating the argument distribution \(q\) on the right-side of (50) by \(p_1\) and \(p_2\), respectively.

**Lemma 3.** For any two pmfs \(p_1, p_2 \in \mathcal{P}(\mathcal{Y})\) with full supports, it holds that
\[
C(p_1, p_2) \leq 2B(p_1, p_2).
\]

**Proof:** The proof follows from an alternative characterization (instead of (2)) of the \(C(p_1, p_2)\) as (cf. [18])
\[
C(p_1, p_2) = \min_{q \in \mathcal{P}(\mathcal{Y})} \max \left( D(q\parallel p_1), D(q\parallel p_2) \right).
\] (54)
and upon noting that the objective function for the optimization problem in (54) is always no larger than that for the one in (50).

**A. Proof of Theorem 1**

Since we consider the error exponent as \(n\) goes to infinity while \(M\) and hence the number of hypotheses is fixed, the ML test, which maximizes the error exponent for the average error probability (averaged over all hypotheses), will also achieve the best error exponent for the maximal error probability. In particular, for any \(y^{Mn} = (y^{(1)}, \ldots, y^{(M)}) \in \mathcal{Y}^{Mn}\), with \(\gamma_{y^{(i)}} = \gamma_i, i = 1, \ldots, M\), conditioned on the \(i\)-th coordinate being the outlier, it follows from (10) that the ML test is
\[
\delta(y^{Mn}) = \arg\min_{i=1,\ldots,M} U_i(y^{Mn}),
\]
where for each \(i = 1, \ldots, M\),
\[
U_i(y^{Mn}) \triangleq D(\gamma_i\parallel \mu) + \sum_{j \neq i} D(\gamma_j\parallel \pi).
\] (55)
By the symmetry of the problem, it is clear that \( P_1 \{ \delta 
eq i \} \) is the same for every \( i = 1, \ldots, M \); hence,

\[
\max_{i=1,\ldots,M} P_1 \{ \delta \neq i \} = P_1 \{ \delta \neq 1 \}.
\]

It now follows from

\[
P_1 \{ \delta \neq 1 \} \leq P_1 (\cup_{j \neq i} \{ U_1 \geq U_j \}),
\]

that

\[
P_1 \{ U_1 \geq U_2 \} \leq P_1 \{ \delta \neq 1 \} \leq \sum_{j=2}^{M} P_1 \{ U_1 \geq U_j \}.
\]

Next, we get from (55) that

\[
P_1 \{ U_1 \geq U_2 \} = P_1 \{ D (\gamma_1 || \mu) + D (\gamma_2 || \pi) \geq D (\gamma_1 || \pi) + D (\gamma_2 || \mu) \}.
\]

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

\[
E = \{(q_1, q_2) : D (q_1 || \mu) + D (q_2 || \pi) \geq D (q_1 || \pi) + D (q_2 || \mu) \},
\]

we get that the exponent for \( P_1 \{ U_1 \geq U_2 \} \) is given by the value of the following optimization problem

\[
\min_{\{q_1, q_2 \in P(Y)\}} \left( D (q_1 || \mu) + D (q_2 || \pi) \right). \tag{58}
\]

Note that the objective function in (58) is convex in \((q_1, q_2)\), and the constraint is linear in \((q_1, q_2)\). It then follows that the optimization problem in (58) is convex. Consequently, strong duality holds for the optimization problem (58) \cite{19}. Then by solving the Lagrangian dual of (58), its solution can be easily computed to be \(2B(\mu, \pi)\).

By the symmetry of the problem, the exponents of \( P_1 \{ U_1 \geq U_i \}, i \neq 1 \), are the same, i.e., for every \( i = 2, \ldots, M \), we get

\[
\lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ U_1 \geq U_i \} = 2B(\mu, \pi). \tag{59}
\]

From (57), (59), using the union bound and that \( \lim_{n \to \infty} \frac{\log M}{n} = 0 \), we get that

\[
\lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ \delta \neq 1 \} = 2B(\mu, \pi). \tag{60}
\]

It is now left to prove that when only \( \pi \) is known, our proposed test in (13), (14) also achieves the error exponent \(2B(\mu, \pi)\). In particular, it follows from the same argument leading to (60) that

\[
\lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ \delta' \neq 1 \} \geq \lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ U_1^{\text{hyp}} \geq U_2^{\text{hyp}} \}. \tag{61}
\]

The exponent on the right-side of (61) can be computed by applying Lemma 1 with \( J = 2, p_1 = \mu, p_2 = \pi \), and (cf.(14))

\[
E = \{(q_1, q_2) : D(q_2 || \pi) \geq D(q_1 || \pi) \}
\]

to be

\[
\min_{\{q_1, q_2 \in P(Y)\}} \left( D (q_1 || \mu) + D (q_2 || \pi) \right). \tag{63}
\]
The optimal value of (63) can be computed as follows

\[
\min_{q_1, q_2 \in \mathcal{P}(Y)} \left( D(q_1 \| \mu) + D(q_2 \| \pi) \right) \geq \min_{q_1} \left( D(q_1 \| \mu) + D(q_1 \| \pi) \right) = 2B(\mu, \pi),
\]

where the inequality in (65) stems from substituting the constraint in (64) into the objective function, and the equality in (66) follows from Lemma 2. Since the minimum in (65) is achieved by \( q_1 = q^* \) in (51) with \( p_1 = \mu, \ p_2 = \pi \), and \( q_1 = q_2 = q^* \) satisfy the constraint in (64), the inequality in (65) is in fact an equality.

B. Proof of Theorem 3

When \( \mu \) and \( \pi \) are unknown, we adopt our universal test \( \delta \) in (15) and (16). In particular, the same argument leading to (69) yields that

\[
\lim_{n \to \infty} - \frac{1}{n} \log \mathbb{P}_1\{\delta \neq 1\} \geq \lim_{n \to \infty} - \frac{1}{n} \log \mathbb{P}_1\{U^{\text{univ}}_1 \leq U^{\text{univ}}_2\}. \tag{67}
\]

The exponent on the right-side of (67) can be computed by applying Lemma 1 with \( J = M \), \( p_1 = \mu \), \( p_j = \pi \), \( j = 2, \ldots, M \), and (cf. (16))

\[
E = \left\{ (q_1, \ldots, q_M) : \sum_{j \neq 1} D(q_j \| \frac{\sum_{k \neq j} q_k}{M-1}) \geq \sum_{j \neq 2} D(q_j \| \frac{\sum_{k \neq j} q_k}{M-1}) \right\} \tag{68}
\]

to be

\[
\min_{(q_1, \ldots, q_M) \in E} \left( D(q_1 \| \mu) + D(q_2 \| \pi) + \ldots + D(q_M \| \pi) \right). \tag{69}
\]

Unlike the convex optimization problems in (58) and (63), the optimization problem in (69) for the completely universal setting is much more complicated, and a closed-form solution is not available. However, we show that the value of (69) is strictly positive for every \( \mu \neq \pi \). In particular, it is not hard to see that the objective function is continuous in \( q_1, \ldots, q_M \) and the constraint set \( E \) is compact. Therefore the minimum in (69) is achieved by some \( (q^*_1, \ldots, q^*_M) \in E \). Note that the objective function in (69) is always nonnegative. In order for the objective function in (69) to be zero, the minimizing \( (q^*_1, \ldots, q^*_M) \) has to satisfy that \( q^*_1 = \mu, q^*_i = \pi \), \( i = 2, \ldots, M \). Since this collection of distributions is not in the constraint set \( E \) in (68), we get that the optimal value of (69) is strictly positive for every \( \mu \neq \pi \).

C. Proof of Theorem 4

By the continuity of the objective function on the right-side of (21) and the compactness of the constraint set (22), for each \( M \geq 3 \), the optimal value on the right-side of (21), denoted by \( V^* \), is achieved by some \( (q^*_1, \ldots, q^*_M) \).

It follows from (21) and (22) that

\[
V^* \geq D(q^*_1 \| \mu) + \sum_{j \neq 1} D(q^*_j \| \pi) - \sum_{j \neq 1} D\left( q^*_j \parallel \frac{\sum_{k \neq j} q^*_k}{M-1} \right) + \sum_{j \neq 2} D\left( q^*_j \parallel \frac{\sum_{k \neq j} q^*_k}{M-1} \right)
\]

\[
= D(q^*_1 \| \mu) + \sum_{j \neq 2} D\left( q^*_j \parallel \frac{\sum_{k \neq j} q^*_k}{M-1} \right) + \sum_{j \neq 1} \sum_{y \in Y} q^*_j(y) \log \left( \frac{1}{M-1} \sum_{k \neq 1} q^*_k(y) \right)
\]
\[ D(q_1^* \| \mu) + \sum_{j \neq 1} D\left(q_j^* \parallel \sum_{k \neq j}^M q_k^* \right) + (M - 1) D\left(\frac{\sum_{k=1}^M q_k^* \parallel \pi}{M-1}\right) \]
\[ \geq D(q_1^* \| \mu) + D\left(q_1^* \parallel \sum_{k \neq 1}^M q_k^* \right) \]
\[ \geq 2B\left(\mu, \frac{q_1^*}{M-1}\right) + \sum_{k \neq 1}^M q_k^* \]
\[ = 2B\left(\mu, \frac{q_1^*}{M-1} + \frac{M-2}{M-1} \left(\sum_{k \neq 1}^{M-1} q_k^* \right)\right) \tag{70} \]

where the last inequality follows Lemma 2.

On the other hand, it follows from (18) that the value on the right-side of (21), \( V^* \), satisfies
\[ 2B(\mu, \pi) \geq V^* = D\left(q_1^* \parallel \mu\right) + \sum_{j \neq 1} D\left(q_j^* \parallel \pi\right) \]
\[ \geq \sum_{j=3}^M D\left(q_j^* \parallel \pi\right) \]
\[ \geq (M - 2) D\left(\frac{1}{M-2} \sum_{k=3}^M q_k^* \parallel \pi\right) \tag{71} \]

where the last inequality follows from the convexity of relative entropy.

Combining (70) and (71), we get that the value \( V^* \) on the right-side of (21) is lower bounded by
\[ \min_{q_1, q \in P(Y)} 2B\left(\mu, \frac{q_1}{M-1}, \frac{M-2}{M-1} q\right) \tag{72} \]

Note that the constraint in (72) can be equally written as
\[ D\left(q_1 \parallel \pi\right) + (M - 2) D\left(q \parallel \pi\right) \leq 2B(\mu, \pi) + D\left(q_1 \parallel \pi\right) \]

Also by the convexity of relative entropy, it follows that
\[ D\left(q_1 \parallel \pi\right) + (M - 2) D\left(q \parallel \pi\right) \geq (M - 1) D\left(\frac{M-1}{M} q_1 + \frac{M-2}{M} q \parallel \pi\right) \]

As a result, the optimal value of (72) is further lower bounded by the optimal value of
\[ \min_{q_1, q \in P(Y)} 2B\left(\mu, \frac{1}{M-1} q_1 + \frac{M-2}{M-1} q\right) \tag{73} \]

By the fact that \( \pi \) has full support, it holds that
\[ D\left(q_1 \parallel \pi\right) \leq -\log \left(\min_{y \in Y} \pi(y)\right) = C_\pi \leq \infty \tag{74} \]

Proceeding from (73), by using (74), we get that the optimal value of (21) is lower bounded by
\[ \min_{q' \in P(Y)} 2B\left(\mu, q'\right) \leq \frac{2B(\mu, \pi) + C_\pi}{D(q' \parallel \pi)} \tag{75} \]
For any $\mu, \pi \in \mathcal{P}(\mathcal{Y})$ with full supports, it holds that
\[
\lim_{M \to \infty} \frac{1}{M - 1} (2B(\mu, \pi) + C_\pi) = 0.
\]
This and the continuity of $D(q\|\pi)$ in $q$ ($\pi$ has a full support) establish (23): the asymptotic efficiency of our test.

Furthermore, for any $\mu, \pi \in \mathcal{P}(\mathcal{Y})$, $\mu \neq \pi$, the value of $\frac{1}{M} (2B(\mu, \pi) + C(\pi))$ is strictly decreasing with $M$. Consequently, the feasible set in (24) is nonincreasing with $M$, and hence the optimal value of (24) is nondecreasing with $M$.

D. Proof of Proposition 4

The proposition follows as a special case of the second assertion of Theorem 11, the proof of which is deferred to Appendix J.

E. Proof of Theorem 5

We start by establishing universal consistency of the test under the null hypothesis. To this end, note that
\[
\mathbb{P}_0 \{ \delta \neq 0 \} \leq \mathbb{P}_0 \left( \bigcup_{j=1}^M \{ U_{j}^{\text{univ}} \geq \lambda_n \} \right)
\leq \sum_{j=1}^M \mathbb{P}_0 \{ U_{j}^{\text{univ}} \geq \lambda_n \}
= M \mathbb{P}_0 \{ U_1^{\text{univ}} \geq \lambda_n \},
\]
where the last equality follows from the fact that all $y_i$, $i = 1, \ldots, M$, are identically distributed according to $\pi$.

We now proceed to bound $\mathbb{P}_0 \{ U_1^{\text{univ}} \geq \lambda_n \}$ as follows:
\[
\mathbb{P}_0 \{ U_1^{\text{univ}} \geq \lambda_n \} = \mathbb{P}_0 \left\{ \sum_{j \neq 1} D \left( \gamma_j \bigg\| \frac{\sum_{k \neq 1} \gamma_k}{M-1} \right) \geq \lambda_n \right\}
= \mathbb{P}_0 \left\{ \sum_{j \neq 1} D (\gamma_j \| \pi) - (M-1) D \left( \frac{\sum_{k \neq 1} \gamma_k}{M-1} \bigg\| \pi \right) \geq \lambda_n \right\}
\leq \mathbb{P}_0 \left\{ \sum_{j \neq 1} D (\gamma_j \| \pi) \geq \lambda_n \right\}
\leq \mathbb{P}_0 \left( \bigcup_{j \neq 1} \left\{ D (\gamma_j \| \pi) \geq \frac{1}{M-1} \lambda_n \right\} \right)
\leq (M-1) \mathbb{P}_0 \left\{ D (\gamma_2 \| \pi) \geq \frac{1}{M-1} \lambda_n \right\},
\]
where the first inequality follows from the non-negativity of the relative entropy, and the last inequality follows from the fact that all $y_i$, $j \neq 1$, are identically distributed according to $\pi$. By the fact that the set of all possible empirical distributions of $(y_1, \ldots, y_n)$ is upper bounded by $(n+1)^{3|\mathcal{Y}|}$ (cf. [16] Theorem 11.1.1), and (4), we get that
\[
\mathbb{P}_0 \left\{ D (\gamma_2 \| \pi) \geq \frac{1}{M-1} \lambda_n \right\} \leq (n+1)^{3|\mathcal{Y}|} \exp \left( - \frac{n}{M-1} \lambda_n \right).
\]
It then follows from (76), (77) and (78) that
\[
\mathbb{P}_0 \{ \delta \neq 0 \} \leq M^2 \exp \left\{ - \frac{n}{M-1} \lambda_n + |\mathcal{Y}| \log(n+1) \right\}. \tag{79}
\]
By choosing \( \lambda_n = 2(M - 1)|\beta|^\log(n+1)/n \), we get from (79) that
\[
\lim_{n \to \infty} P_0(\delta \neq 0) = 0.
\]

Next we treat the exponent for the conditional probability of error under every non-null hypothesis. In particular, by the symmetry of the test (26) among all the \( M \) non-null hypotheses, it suffices to consider the conditional error probability under just the first hypothesis, which can be upper bounded as follows:
\[
\mathbb{P}_1 \{ \delta \neq 1 \} \leq \mathbb{P}_1 \left( \bigcup_{j \neq 1} \left\{ U_1^{\text{univ}} \geq U_j^{\text{univ}} - \lambda_n \right\} \right) \\
\leq \sum_{j \neq 1} \mathbb{P}_1 \left\{ U_1^{\text{univ}} \geq U_j^{\text{univ}} - \lambda_n \right\} \\
\leq (M - 1) \mathbb{P}_1 \left\{ U_1^{\text{univ}} \geq U_2^{\text{univ}} - \lambda_n \right\}. 
\]

(80)

For an arbitrary \( \lambda_0 > 0 \), as \( \lambda_n \to 0 \), it holds that \( \lambda_n \leq \lambda_0 \) for \( n \) sufficiently large and hence that
\[
\mathbb{P}_1 \left\{ U_1^{\text{univ}} \geq U_2^{\text{univ}} - \lambda_0 \right\} \leq \mathbb{P}_1 \left\{ U_1^{\text{univ}} \geq U_2^{\text{univ}} - \lambda_0 \right\}. 
\]

(81)

The exponent of the right-side of (81) can be computed by applying Lemma 1 with \( J = M \), \( p_1 = \mu \), \( p_j = \pi \), \( j = 2, \ldots, M \) and (cf.(16))
\[
E(\lambda_0) \triangleq \left\{ (q_1, \ldots, q_M) : \sum_{j \neq 1} D(q_j \mid \sum_{k \neq j} q_k / M - 1) \geq \sum_{j \neq 2} D(q_j \mid \sum_{k \neq j} q_k / M - 1) - \lambda_0 \right\}
\]
to be
\[
\min_{(q_1, \ldots, q_M) \in E(\lambda_0)} D(q_1 \mid \mu) + D(q_2 \mid \pi) + \ldots + D(q_M \mid \pi).
\]
(82)

Since \( \lambda_0 \) can be arbitrarily close to zero, the exponent for the left-side of (81) is lower bounded by
\[
\lim_{\lambda_0 \to 0} \min_{(q_1, \ldots, q_M) \in E(\lambda_0)} D(q_1 \mid \mu) + D(q_2 \mid \pi) + \ldots + D(q_M \mid \pi).
\]

Let
\[
E \triangleq \left\{ (q_1, \ldots, q_M) : \sum_{j \neq 1} D(q_j \mid \sum_{k \neq j} q_k / M - 1) \geq \sum_{j \neq 2} D(q_j \mid \sum_{k \neq j} q_k / M - 1) \right\}.
\]

By the fact that \( E(\lambda_0) \) is closed and compact for any \( \lambda_0 > 0 \), and that the objective function in (82) is continuous, the exponent for the left-side of (81) is lower bounded by
\[
\min_{(q_1, \ldots, q_M) \in E} D(q_1 \mid \mu) + D(q_2 \mid \pi) + \ldots + D(q_M \mid \pi),
\]
(83)
as required.

F. Proof of Proposition 7

The proposition follows from a well-known result in detection and estimation in the context of multihypothesis testing problem [20]. In particular, the optimal error exponent for testing \( M \) hypotheses with i.i.d. observations with respect to \( p_1, p_2, \ldots, p_M \) is characterized as \( \min_{1 \leq i < j \leq M} C(p_i, p_j) \).

When all the \( \{\mu_i\}_{i=1}^M \) and \( \pi \) are known, the underlying outlier hypothesis testing problem is just a multihypothesis testing problem based on i.i.d. vector observations (with \( M \) independent components) and consequently, the optimal
error exponent can be computed as
\[
\min_{S \neq S'} C \left( \prod_{i \in S} \mu_i(y_i) \prod_{j \notin S} \pi(y_j), \prod_{i \in S'} \mu_i(y_i) \prod_{j \notin S'} \pi(y_j) \right)
\]
\[
= \min_{S \neq S'} C \left( \prod_{i \in S \setminus S'} \mu_i(y_i) \prod_{j \in S' \setminus S} \pi(y_j), \prod_{i \in S' \setminus S} \mu_i(y_i) \prod_{j \notin S' \setminus S} \pi(y_j) \right)
\]
\[
= \min_{S \neq S'} \max_{s \in [0,1]} - \log \left( \sum_{y_i, y_j} \prod_{i \in S \setminus S'} \mu_i(y_i)^s \pi(y_i)^{1-s} \prod_{j \in S' \setminus S} \pi(y_j)^s \mu_j(y_j)^{1-s} \right) \tag{84}
\]
\[
= \min_{1 \leq i < j \leq M} \max_{s \in [0,1]} - \log \left( \sum_{y_i, y_j} (\mu_i(y_i)^s \pi(y_i)^{1-s})(\pi(y_j)^s \mu_j(y_j)^{1-s}) \right) \tag{85}
\]
\[
= \min_{1 \leq i < j \leq M} C(\mu_i(y) \pi(y'), \pi(y) \mu_j(y')),
\]
where the equality in (85) follows by virtue of the fact that the outer minimum in (84) is attained among the pairs of \(S, S'\), with the largest number of coordinates in their intersections: \(T - 1\).

When all the outlier coordinates are identically distributed, i.e., \(\mu_i = \mu, i = 1, \ldots, M\), this optimal error exponent can be further simplified to be
\[
\min_{1 \leq i < j \leq M} C(\mu_i(y) \pi(y'), \pi(y) \mu_j(y')) = C(\mu(y) \pi(y'), \pi(y) \mu(y')) = 2B(\mu, \pi). \tag{86}
\]

**G. Proof of Theorem 8**

Let \(\delta\) denote the test specified by (31) and (33) wherein \(\gamma_i, i = 1, \ldots, M\), denote the empirical distributions of observations in the \(M\) coordinates respectively. It follows from the fact that for every \(S \subseteq S\),

\[
P_S \{\delta \neq S\} \leq P_S \left( \bigcup_{S' \neq S} \{U_S^{yp} \geq U_{S'}^{yp}\} \right)
\]

that
\[
\max_{S \neq S'} P_S \{U_S^{yp} \geq U_{S'}^{yp}\} \leq \max_{S \neq S} P_S \{\delta \neq S\} \leq \max_{S \neq S} P_S \{U_S^{yp} \geq U_{S'}^{yp}\} \leq (|S| - 1) \max_{S \neq S} P_S \{U_S^{yp} \geq U_{S'}^{yp}\}. \tag{87}
\]

Next, we get from (31) that for any \(S \neq S' \subseteq S\),

\[
P_S \{U_S^{yp} \geq U_{S'}^{yp}\} = P_S \left( \sum_{i \notin S} D(\gamma_i || \pi) \geq \sum_{i \notin S'} D(\gamma_i || \pi) \right).
\]

Applying Lemma 1 with \(J = M\), \(p_i = \mu_i, i \in S\), \(p_j = \pi, j \notin S\),
\[
E = \left\{ (q_1, \ldots, q_M) : \sum_{i \notin S} D(q_i || \pi) \geq \sum_{i \notin S'} D(q_i || \pi) \right\}, \tag{88}
\]
we get that the exponent for \( P_S \{ U^\text{typ}_S \geq U^\text{typ}_{S'} \} \) is given by the value of the following optimization problem

\[
\min_{\{q_i\}_{i \in S \setminus S'}, \{q_j\}_{j \in S', S}} \sum_{i \in S \setminus S'} D(q_i \| \mu_i) + \sum_{j \in S' \setminus S} D(q_j \| \pi). \tag{89}
\]

We now show that the optimum value in \((89)\) is equal to \( \sum_{i \in S \setminus S'} 2B(\mu_i, \pi) \). First, we show that the latter is a lower bound for the former. Substituting the constraint in \((89)\) into the objective function, we get that the value of \((89)\) is lower bounded by

\[
\min_{\{q_i\}_{i \in S \setminus S'}, \{q_j\}_{j \in S', S}} D(q_i \| \mu_i) + D(q_i \| \pi) = \sum_{i \in S \setminus S'} 2B(\mu_i, \pi), \tag{90}
\]

where the equality follows from Lemma 2. Second, note that \(|S' \setminus S|\) is always equal to \(|S \setminus S'|\), and, hence, we can make a suitable correspondence between elements of \(S \setminus S'\) to those of \(S' \setminus S\). The converse implication now follows by assigning for every \(i \in S' \setminus S\), and the corresponding \(j \in S' \setminus S\), \( q_i = q_j = \frac{\mu_i(y)^{1/2} \pi(y)^{1/2}}{\sum_{y} \mu_i(y)^{1/2} \pi(y)^{1/2}} \), and note that this assignment trivially satisfies the constraint in \((89)\) and gives rise to the objective function being equal to \( \sum_{i \in S \setminus S'} 2B(q_i, \pi) \).

Lastly, it follows from \((87)\) that

\[
\lim_{n \to \infty} -\frac{1}{n} \log \left( \max_{S \in \mathcal{S}} P_S \{ \delta \neq S \} \right) = \min_{S \neq S'} \sum_{i \in S \setminus S'} 2B(\mu_i, \pi) = \min_{1 \leq i \leq M} 2B(\mu_i, \pi).
\]

When \(\mu_i = \mu, i = 1, \ldots, M\),

\[
\min_{1 \leq i \leq M} 2B(\mu_i, \pi) = 2B(\mu, \pi).
\]

H. Proof of Theorem 2

Consider the test \(\delta\) specified by \((32)\) and \((34)\) wherein \(\gamma_i, i = 1, \ldots, M\), denote the empirical distributions of observations in the \(M\) coordinates respectively. It now follows in the manner similar to \((87)\) that

\[
\max_{S \neq S'} P_S \{ U^\text{univ}_S \geq U^\text{univ}_{S'} \} \leq \max_{S \in \mathcal{S}} P_S \{ \delta \neq S \} \leq \max_{S \in \mathcal{S}} \sum_{S' \neq S} P_S \{ U^\text{univ}_S \geq U^\text{univ}_{S'} \} \leq (|S| - 1) \max_{S \neq S'} P_S \{ U^\text{univ}_S \geq U^\text{univ}_{S'} \}. \tag{91}
\]

The assertion \((40)\) now follows from \((91)\) upon noting that the application of Lemma 1 with \(J = M\), \(p_i = \mu_i, i \in S\), \(p_j = \pi, j \notin S\), and

\[
E = \left\{ (q_1, \ldots, q_M) : \sum_{i \notin S} D \left( q_i \left\| \frac{\sum_{k \notin S, k \neq i} q_k}{M - |S|} \right. \right) \geq \sum_{i \notin S'} D \left( q_i \left\| \frac{\sum_{k \notin S', k \neq i} q_k}{M - |S'|} \right. \right), \tag{92} \right\}
\]

gives that the exponent for \( P_S \{ U^\text{univ}_S \geq U^\text{univ}_{S'} \} \) is equal to

\[
\sum_{i \notin S} D \left( q_i \left\| \frac{\sum_{k \notin S, k \neq i} q_k}{M - |S|} \right. \right) \geq \sum_{i \notin S'} D \left( q_i \left\| \frac{\sum_{k \notin S', k \neq i} q_k}{M - |S'|} \right. \right) \sum_{i \in S} D \left( q_i \left\| \mu_i \right. \right) + \sum_{j \notin S} D \left( q_j \left\| \pi \right. \right). \tag{93}
\]

Lastly, the assertion of universally exponential consistency of the proposed test follows from the compactness of the the feasible set of \((92)\), continuity of the objective function in \((93)\), and the fact that the objective function of \((93)\) can only be zero at a collection \((q_i = \mu_i, i \in S, q_j = \pi, j \notin S)\), which is not in the constraint set.
I. Proof of Theorem 11

First let denote the minimizing $S$ and $S'$ in the outer minimum of (40) by $S^*$ and $S'^*$ respectively, and the minimizing tuple $q_1,\ldots,q_M$ in the inner minimum of (40) by $q_1^*,\ldots,q_M^*$. Then, we get that the achievable error exponent in (40) is lower bounded as
\[
\begin{align*}
\geq & \sum_{i \in S^*} D\left(\mu_i \| \mu_i\right) + \sum_{j \notin S^*} D\left(q_j^* \| \pi\right) - \sum_{j \notin S'^*} D\left(\frac{\sum_{i \in S^*} \mu_i q_i^*}{M-T}\| \pi\right) + \sum_{j \notin S'^*} D\left(q_j^* \| \frac{\sum_{i \in S'^*} \mu_i q_i^*}{M-T}\right) \\
= & \sum_{i \in S^*} D\left(q_j^* \| \mu_i\right) + \sum_{j \notin S'^*} D\left(q_j^* \| \frac{\sum_{i \in S'^*} \mu_i q_i^*}{M-T}\right) + (M-T) D\left(\frac{\sum_{i \in S'^*} \mu_i q_i^*}{M-T}\| \pi\right) \\
\geq & D\left(q_j^* \| \mu_i\right) + D\left(q_j^* \| \frac{\sum_{i \in S'^*} \mu_i q_i^*}{M-T}\right) \\
\geq & 2B\left(\mu_i, \frac{\sum_{i \in S'^*} \mu_i q_i^*}{M-T}\right),
\end{align*}
\]
where $t$ is an arbitrarily chosen element in $S^*\setminus S'^*$.

On the other hand, it follows from Proposition 7 that
\[
\begin{align*}
\min_{1 \leq i < j \leq M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) & \geq \sum_{i \in S^*} D\left(q_j^* \| \mu_i\right) + \sum_{j \notin S^*} D\left(q_j^* \| \pi\right) \\
& \geq \sum_{j \notin S'^* \cup S'^*} D\left(q_j^* \| \pi\right) \\
& \geq (M-T-|S^*\setminus S'^*|) D\left(\frac{\sum_{i \in S^* \setminus S'^*} \mu_i q_i^*}{(M-T-|S^*\setminus S'^*|)} \| \pi\right).
\end{align*}
\]
It now follows from (95) that
\[
\begin{align*}
(M-T) D\left(\frac{\sum_{i \in S^* \setminus S'^*} \mu_i q_i^*}{M-T} \| \pi\right) & \leq (M-T-|S^*\setminus S'^*|) D\left(\frac{\sum_{i \in S^* \setminus S'^*} \mu_i q_i^*}{(M-T-|S^*\setminus S'^*|)} \| \pi\right) \\
& + (|S^*\setminus S'^*|) D\left(\frac{\sum_{i \in S^* \setminus S'^*} \mu_i q_i^*}{|S^*\setminus S'^*|} \| \pi\right) \\
& \leq \min_{1 \leq i < j \leq M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) + |S^*\setminus S'^*|C_\pi \\
& \leq \min_{1 \leq i < j \leq M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) + TC_\pi.
\end{align*}
\]
The lower bound in (44) now follows from (94) and (96).

The assertion (45) now follows from (44), Proposition 7 and the continuity of $B(\mu,q)$ and $D(q\|\pi)$ in the argument $q$. The assertion (46) follows as a special case of (45).

It is now left only to prove the universally asymptotically exponential consistency of the proposed test in (32), (34). Having proved (45), this assertion now follows upon noting that for every $i,j$, $1 \leq i < j \leq M$, it holds that
\[
C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) \leq 2B(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')) \\
= -2\log\left(\sum_{y,y' \in Y \times Y} (\mu_i(y)\pi(y'))^{\frac{1}{2}} (\pi(y)\mu_j(y'))^{\frac{1}{2}} \right) \\
= 2B(\mu_i, \pi) + 2B(\mu_j, \pi),
\]
where the first inequality above follows from Lemma 3.

J. Proof of Theorem 17

We first prove that for every hypothesis set excluding the null hypothesis, our proposed test in (47), (48) is universally exponentially consistent.
Following the same argument leading to (87), it suffices to show that for any $S, S' \in S, S' \neq S$,

$$\lim_{n \to \infty} -\frac{1}{n} \log \left( \mathbb{P}_S \left\{ \hat{U}^{\text{univ}}_S \geq \hat{U}^{\text{univ}}_{S'} \right\} \right) > 0.$$ (97)

Applying Lemma 1 with $J = M$, $p_i = \mu$, $i \in S$, $p_j = \pi$, $j \notin S$, and

$$E_{(S, S')} = \left\{ (q_1, \ldots, q_M) : \sum_{i \in S} D \left( q_i \left\| \frac{\sum_{k \in S} q_k}{T} \right\| \right) + \sum_{j \notin S} D \left( q_j \left\| \frac{\sum_{k \in S} q_k}{T} \right\| \right) \right\},$$

we get that the exponent for $\mathbb{P}_S \left\{ \hat{U}^{\text{univ}}_S \geq \hat{U}^{\text{univ}}_{S'} \right\}$ is given by the value of the following optimization problem

$$\min_{\{q_1, q_2, \ldots, q_M\} \in E_{(S, S')}} \sum_{i \in S} D(q_i \| \mu) + \sum_{j \notin S} D(q_j \| \pi).$$ (98)

The solution to $\sum_{i \in S} D(q_i \| \mu) + \sum_{j \notin S} D(q_j \| \pi) = 0$ is uniquely given by $q_i = \mu$ for $i \in S$, $q_j = \pi$ for $j \notin S$. Because $|S| < M/2$, $|S'| < M/2$, there is no $S, S' \in S$, such that $S = \{1, 2, \ldots, M\} \setminus S'$. Let $q_i = \mu$ for $i \in S$, $q_j = \pi$ for $j \notin S$, it then follows that

$$0 = \sum_{i \in S} D \left( q_i \left\| \frac{\sum_{k \in S} q_k}{T} \right\| \right) + \sum_{j \notin S} D \left( q_j \left\| \frac{\sum_{k \in S} q_k}{T} \right\| \right) < \sum_{i \in S} D \left( q_i \left\| \frac{\sum_{k \in S} q_k}{T} \right\| \right) + \sum_{j \notin S} D \left( q_j \left\| \frac{\sum_{k \in S'} q_k}{T} \right\| \right)$$

for any $S, S' \in S, S' \neq S$. In other words, the objective function in (98) is strictly positive at any feasible $(q_1, q_2, \ldots, q_M)$. By the continuity of the objective function in (98) and the fact that $E_{(S, S')}$ is compact for any $S, S' \in S$, it holds that the value of the optimization function in (98) is strictly positive for every pair of $S, S' \in S, S \neq S'$, which establishes the universally exponential consistency of our proposed universal test.

Next to prove the second assertion, it suffices to prove that even when the typical distribution is known, there cannot exist a universally exponentially consistent test in differentiating the null hypothesis from any other hypothesis with a positive number of outliers. To this end, let $S \subset \{1, 2, \ldots, M\}$, $|S| \geq 1$ denote an arbitrary set of outlier coordinates. To distinguish between the null hypothesis and $S$, a test is done based on a decision rule $\delta : \mathcal{Y}^{\text{Mn}} \to \{0, 1\}$, where $0$ corresponds to the null hypothesis and $1$ the hypothesis with $S$ being the outliers. It should be noted that $\delta$ can only be a function of $\pi$ and the observations $\mathcal{Y}^{\text{Mn}}$.

We first show that in order to distinguish between the null hypothesis and $S$, the empirical distributions of all the coordinates $\gamma_1, \ldots, \gamma_M$ and $\pi$ are sufficient statistics for the error exponent. In particular, we now show that given any test, there is another test that achieves the same error exponent with its decision being made based only on the empirical distributions of all $M$ coordinates and $\pi$. To this end, for feasible empirical distributions (for certain $n$) $\gamma_1, \ldots, \gamma_M$, let us denote the set of all vector observations conforming to these empirical distributions by $\mathcal{T}_{(\gamma_1, \ldots, \gamma_M)}$. Among these vector observations, let us denote the set of vector observations for which $\delta$ decides for the null hypothesis by $\mathcal{T}_{(\gamma_1, \ldots, \gamma_M)}$, which may depend on $\pi$. Now consider another test $\delta'$ which decides on one of the two hypotheses based only on $\gamma_1, \ldots, \gamma_M$ and $\pi$. Specifically, this new test is such that for all vector observations with empirical distributions $\gamma_1, \ldots, \gamma_M$, it decides for the null hypothesis if $|\mathcal{T}_{(\gamma_1, \ldots, \gamma_M)}| \geq \frac{1}{2}|\mathcal{T}_{(\gamma_1, \ldots, \gamma_M)}|$, and for $S$ otherwise. It follows from this construction of $\delta'$ that for any $\mu$ and $\pi$,

$$\max \left\{ \mathbb{P}_0 \{ \delta' \neq 0 \}, \mathbb{P}_1 \{ \delta' \neq 1 \} \right\} \leq 2 \max \left\{ \mathbb{P}_0 \{ \delta \neq 0 \}, \mathbb{P}_1 \{ \delta \neq 1 \} \right\},$$

where $\mathbb{P}_0, \mathbb{P}_1$ are the distributions under the null hypothesis, and under the hypothesis with $S$ being the outliers, respectively. Consequently, the error exponent achievable by $\delta'$ is the same as that achievable by $\delta$ for any $\mu, \pi, \mu \neq \pi$. 
Having shown that the empirical distributions of the $M$ coordinates and $\pi$ are sufficient statistics, it suffices to consider tests that depend only on $\gamma_1, \ldots, \gamma_M$, and $\pi$. In particular, for any such $\epsilon$, let assume that for any $\pi$, there exists $\epsilon = \epsilon(\pi) > 0$ such that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_0 \{ \delta \neq 0 \} > \epsilon. \quad (99)$$

It now follows from (99) and Lemma 1 that the set $A$ of all $M$ tuples $(y^{(1)}, \ldots, y^{(M)}) \in Y^M$, whose empirical distributions $(\gamma_1, \ldots, \gamma_M)$ lie in the following set

$$E \triangleq \left\{ (q_1, \ldots, q_M) : \sum_{i \in S} D(q_i \| \pi) + \sum_{j \notin S} D(q_j \| \pi) \leq \frac{\epsilon}{2} \right\},$$

must satisfy

$$A \subseteq \{ \delta = 0 \}. \quad (100)$$

By applying Lemma 1 again, but now with respect to the hypothesis with $S$ being the outliers, we get that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ \delta \neq 1 \} \leq \min_{(q_1, \ldots, q_M) \in E} \sum_{i \in S} D(q_i \| \mu) + \sum_{j \notin S} D(q_j \| \mu). \quad (101)$$

Since $\epsilon$ is independent of $\mu$, and $\mu$ can be chosen arbitrarily close to $\pi$, we can pick $\mu$ to be such that $\sum_{i \in S} D(\mu \| \pi) < \frac{\epsilon}{2}$. It now follows from the definition of $E$, (101) and Lemma 1 that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ \delta \neq 1 \} = 0,$$

which establishes the assertion, as had (99) not been held, the error exponent for $\delta$ is also zero.

**K. Proof of Theorem 2**

Without loss of generality, we can consider the following two hypotheses. The first hypothesis has $S_1$ as the set of outliers, and the second hypothesis has $S_2$, where $S_1 \subset S_2$. It suffices to prove that even when $\pi$ and $\{\mu_i\}_{i \in S_1}$ are known, there cannot exist a universally exponentially consistent test in differentiating such two hypotheses. By the same argument as in the proof of Theorem 1, we can consider tests that depend only on the empirical distributions of all the coordinates $\gamma_1, \ldots, \gamma_M$, $\pi$ and $\{\mu_i\}_{i \in S_1}$. Such a test is based on a decision rule $\delta : Y^M \to \{1, 2\}$, where 1 corresponds to the hypothesis with $S_1$ being the outliers and 2 to the hypothesis with $S_2$. In particular, let assume that for any fixed $\pi$ and $\{\mu_i\}_{i \in S_1}$, there exists $\epsilon = \epsilon(\pi, \{\mu_i\}_{i \in S_1}) > 0$ such that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_1 \{ \delta \neq 1 \} > \epsilon, \quad (102)$$

where $P_1$ is the distribution under the hypothesis with $S_1$ being the outliers. It now follows from (102) and Lemma 1 that the set $A$ of all $M$ tuples $(y^{(1)}, \ldots, y^{(M)}) \in Y^M$ whose empirical distributions $(\gamma_1, \ldots, \gamma_M)$ lie in the following set

$$E \triangleq \left\{ (q_1, \ldots, q_M) : \sum_{i \in S_1} D(q_i \| \mu_i) + \sum_{j \notin S_1} D(q_j \| \pi) \leq \frac{\epsilon}{2} \right\} \quad (103)$$

must be such that

$$A \subseteq \{ \delta = 1 \}. \quad (104)$$
By applying Lemma[1] again, but now with respect to the hypothesis with $S_2$ being the outliers, we get that
\[
\lim_{n \to \infty} -\frac{1}{n} \log P_2 \{ \delta \neq 2 \} \leq \min_{(q_1, \ldots, q_M) \in E} \sum_i D(q_i \Vert \mu_i) + \sum_{j \in S_2 \setminus S_1} D(q_j \Vert \mu_j) + \sum_{k \in S_2} D(q_k \Vert \pi),
\]
where $P_2$ is the distribution under the hypothesis with $S_2$ being the outliers. Since $\epsilon$ is independent of $\{ \mu_j \}_{j \in S_2 \setminus S_1}$, we can pick $\{ \mu_j \}_{j \in S_2 \setminus S_1}$ to be such that $\sum_{j \in S_2 \setminus S_1} D(\mu_j \Vert \pi) < \frac{\epsilon}{2}$. It now follows from the definition of $E$, (105) and Lemma[1] that
\[
\lim_{n \to \infty} -\frac{1}{n} \log P_2 \{ \delta \neq 2 \} = 0,
\]
which establishes the assertion.

L. Optimal Test for One Outlier When Only $\mu$ Is Known

Now we address the issue raised in Remark[1]. In particular, when only $\mu$ is known, instead of using the corresponding version of the test proposed in Section III-A1, we adopt the following test $\hat{\delta}$:
\[
\hat{\delta}(y^M) = \arg \min_{i = 1, \ldots, M} D(\gamma_i \Vert \mu),
\]
where $\gamma_i$, denotes the empirical distribution of $y^{(i)}$, $i = 1, \ldots, M$, and the ties in (106) are broken arbitrarily.

It now follows using (106), that
\[
P_1 \{ \hat{\delta} \neq 1 \} \leq (M - 1)P_1 \{ D(\gamma_1 \Vert \mu) \geq D(\gamma_2 \Vert \mu) \}
\]
Applying Lemma[1] with $J = 2$, $p_1 = \mu$, $p_2 = \pi$,
\[
E = \{ (q_1, q_2) : D(q_1 \Vert \mu) \geq D(q_2 \Vert \mu) \},
\]
we get that the exponent for $P_1 \{ \hat{\delta} \neq 1 \}$ is given by the value of the following optimization problem
\[
\min_{q_1, q_2 \in P(Y)} D(q_1 \Vert \mu) + D(q_2 \Vert \pi) \quad D(q_1 \Vert \mu) \geq D(q_2 \Vert \mu)
\geq \min_{q_2} D(q_2 \Vert \mu) + D(q_2 \Vert \pi),
\]
where the inequality follows by substituting the constraint into the objective function and the equality follows from Lemma[2].

REFERENCES


