Second-Order Asymptotically Optimal Universal Outlier Hypothesis Testing

Lin Zhou, Yun Wei and Alfred Hero

Abstract

We revisit the universal outlier hypothesis testing (Li et al., TIT 2014) and derive fundamental limits for the optimal test. In outlying hypothesis testing, one is given multiple observed sequences, where most sequences are generated i.i.d. from a nominal distribution. The task is to discern the set of outlying sequences that are generated according to anomalous distributions. The nominal and anomalous distributions are unknown. We study the tradeoff among the probabilities of misclassification error, false alarm and false reject for tests that satisfy weak conditions on the rate of decrease of these error probabilities as a function of sequence length. Specifically, we propose a threshold-based universal test that ensures exponential decay of misclassification error and false alarm probabilities. We study two constraints on the false reject probabilities, one is that it be a non-vanishing constant and the other is that it have an exponential decay rate. For both cases, we characterize bounds on the false reject probability, as a function of the threshold, for each pair of nominal and anomalous distributions and demonstrate the optimality of our test in the generalized Neyman-Pearson sense. We first consider the case of at most one outlier and then generalize our results to the case of multiple outliers where the number of outliers is unknown and each outlier can follow a different anomalous distribution.

Index Terms

Second-order asymptotic approximations, Error exponents, Misclassification error, False alarm, False reject,

I. INTRODUCTION

In the outlier hypothesis testing (OHT) problem, one observes a number of sequences. The majority of the sequences are i.i.d. samples from a nominal distribution and the rest of the sequences are i.i.d. samples from anomalous distributions different from the nominal distribution. The task in the universal OHT problem is to design a test to discern the set of outlying sequences with high probability when both nominal and anomalous distributions are unknown. Motivated by practical applications in anomaly detection [1], we revisit the OHT problem studied in [2] when the outlier might not be present and derive the performance tradeoff between the probabilities of misclassification error, false alarm and false reject for universal tests. Furthermore, we show that such tests are optimal in the generalized Neyman-Pearson sense for both a second-order asymptotic regime and a large deviations regime. Our second-order asymptotic result provides an approximation to the finite sample performance of the tests. Throughout the paper, we consider the case where the sequences have a finite alphabet.

We first consider the case when there is at most one outlier. Under this setting, the null hypothesis is that there is no outlying sequence while a non-null hypothesis specifies the index of the outlying sequence. Li et al. [2, Theorem 5] showed that the error probability under non-null hypothesis decays exponentially fast and that the error probability under null hypothesis vanishes as the length of observed sequence tends to infinity for the threshold based generalized likelihood ratio test [2, Eq. (25)]. Furthermore, the authors of [2] showed the optimality of their test when the number of observed sequences tends to infinity. A natural question arises: whether or not it is possible to claim optimality for a universal test when the number of observed sequences is finite and when the length of the observed sequences is non-asymptotic. Our first contribution sheds lights on the positive answer for this question. To do so, we decompose the error probability under the non-null hypothesis into the misclassification error probability and the false reject probability, where the false reject event corresponds to falsely claiming that no outlier exists and the misclassification error event corresponds to falsely claiming that a nominal sequence is an outlier. The error probability under the null hypothesis is denoted the probability of false alarm, which is the probability of falsely claiming that an observed sequence is an outlier when no outlier is present. We show that a test, inspired by sequence classification with empirical statistics [3], [4], is optimal under certain conditions, from a second-order or a first-order asymptotic perspective.

We then generalize our results to the case where the number of outliers is unknown and each outlying sequence can be generated from a potentially different anomalous distribution. When the number of outliers is known, Li et al. [2, Theorem 10] derived an achievability decay rate of the error probabilities under each hypothesis and showed asymptotic optimality of their result when the number of the sequences tends to infinity, when the lengths of sequences tend to infinity and when all the outlying sequences are generated from the same anomalous distribution. Furthermore, when the number of outliers is unknown

This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1804800, by a start up grant at the Beihang university, by ARO grant W911NF-15-1-0479 and by DOE grant DE-NA0003921.

Lin Zhou is with the School of Cyber Science and Technology, Beihang University, Beijing, China, 100083 (Email: lzhou@buaa.edu.cn). He was with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, 48109.

Yun Wei is with the Department of Statistical Science, Duke University, Durham, NC 27708 (Email: yun.wei@duke.edu).

Alfred Hero is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, 48109 (Email: hero@eecs.umich.edu).
and when each outlier is generated from the same anomalous distribution, Li et al. [2, Theorem 10] showed that when the null hypothesis is not taken into account, a generalized likelihood ratio test is exponentially consistent. However, the authors of [2] did not provide characterizations of the exponent. One might wonder whether it is possible to provide characterizations of the performance of a universal test where the number of outliers is unknown and where each outlier can be generated from a different anomalous distribution. Our second contribution provides a positive answer to this question and also demonstrates the optimality of the universal test in the generalized Neyman-Pearson sense.

A. Main Contributions

Our main contribution is an analysis of the tradeoff among probabilities of misclassification error, false reject and false alarm probabilities for universal tests in the outlier hypothesis testing problem. For the case with at most one outlier, our results complement [2, Theorem 5] by proposing a new threshold-based universal test, providing a second-order asymptotic approximation to the performance of the universal test with finite length sequences and relaxing the conditions for which the optimality of a universal test can be claimed using a weaker condition inspired by statistical classification [3]. Furthermore, asymptotically, our results in Theorem 3 complement [2, Proposition 4] by identifying a sufficient condition on the pair of nominal and anomalous distributions under which the universal test ensures exponential decay of all three kinds of error probabilities. Finally, the information theoretical quantity that characterizes the performance tradeoff is shown to be a generalized Jensen-Shannon divergence, which is significantly different from the constrained sum of KL divergences in [2, Eq. (26)] or the Bhattacharyya distance in [2, Corollary 6]. For the case admitting multiple outliers, we analyze the performance of a universal test ignorant of the number of outliers where each outlier can be drawn from a different distribution. We note that our results close an existing gap in the theory developed in [2, Section IV], providing explicit equations for the asymptotic performance of the outlier test.

For the case with at most one outlier, we propose a threshold-based universal test that ensures the exponential decay of misclassification error and false alarm probabilities, called the homogeneous error exponent, which simultaneously upper bounds the false reject probabilities as a function of the threshold for any pair of nominal and anomalous distributions. We first derive a second-order asymptotic result that provides an approximation to the performance of the test when the length of each observed sequence is finite. In particular, we show that if the threshold of the test is upper bounded by a certain function of the number of observed sequences, the length of each sequence, and the nominal and anomalous distributions \((P_N, P_A)\), the false reject probability is essentially upper bounded by a constant \(\varepsilon \in (0, 1)\), say \(10^{-3}\). Furthermore, among all tests that can ensure an exponential decay of misclassification error and false alarm probabilities for all pairs of nominal and anomalous distributions, our proposed test has the same homogeneous error exponent and the smallest non-vanishing false reject probability under any pair of nominal and anomalous distributions. This way, the optimality in the above generalized Neyman-Pearson sense of our universal test is ensured for any finite number of observed sequences \(M\) (see [4] for a similar result in the context of statistical classification).

In anomaly detection problems, it may be necessary to maintain a vanishingly small false reject probability when the length of each observed sequence becomes large. To resolve this problem, asymptotically when the lengths of the observed sequences tend to infinity, we derive the exponential decay rate of the false reject probability as a function of the threshold in the universal test. We show, in particular, that the homogeneous error exponent is the threshold of the test for any pair of nominal and anomalous distributions. This way, we establish that, as long as the nominal and anomalous distribution is far in certain distance measure, the universal test is exponentially consistent as all three kinds of error probabilities decay to zero exponentially fast. Conversely, we show that among all tests that can ensure the same speed of exponential decay of misclassification error and false alarm probabilities for all pairs of nominal and anomalous distributions, our proposed test guarantees the largest exponential decay rate for the false reject probability regardless the pair of nominal and anomalous distributions.

B. Related Works

The most closely related work to ours is that of [2], where the authors formulated the outlier hypothesis testing problem, and derived optimal results under constraints on the number of observed sequences, the length of observed sequences and the number of anomalous distributions. Other related work on outlier hypothesis testing is the following. A low complexity test for outlier hypothesis testing was proposed and analyzed in [5]. A distribution free test based on maximum mean discrepancy was proposed in [6] and shown to be exponentially consistent when the number of outliers is known under a certain condition on the number of observed sequences and the length of each sequence. Readers may also refer to [7] for a comprehensive survey of the commonly made assumptions on distributions, definitions of outliers, types of tests, applications and results in outlying sequence detection. Furthermore, the results of [2] were generalized to a sequential setting in [8] where each sequence is observed symbol by symbol until the test is confident enough to make a decision. In [9], the authors studied the quickest outlier detection problem where outlying sequences follow an anomalous distribution after a certain unknown change time and universal tests were proposed to identify the outliers. Finally, in [10] for detecting an outlier from \(M\) sequence streams, the authors studied a special case of the sequential outlier detection problem where at each time only a subset of all sequence symbols can be observed.
Since our proof technique is inspired by statistical classification, we also mention a few works in this domain. In [3], the author studied a binary sequence classification problem and showed that a certain test using empirical distributions of training and testing data is asymptotically optimal with exponentially decreasing misclassification probabilities. The result in [3] was generalized to classification of multiple testing sequences in [11] and to distributed detection in [12]. Finally, a finite sample analysis for the setting of [3] was provided in [4].

C. Organization for the Rest of the Paper

The rest of the paper is organized as follows. In Section II, we set up the notation, formulate the universal outlying sequence detection problem with at most one outlying sequence, propose fundamental limits and present our main results concerning the performance of optimal universal tests. In Section III, we generalize our results to the case of multiple outlying sequences where the number of outlying sequences is known and each outlying sequence is generated according to a potentially different anomalous distribution. Finally, we conclude the paper and discuss future research directions in Section IV. The proofs of all theorems are deferred to appendices.

II. CASE OF AT MOST ONE OUTLYING SEQUENCE

Notation

Random variables and their realizations are in upper (e.g., $\mathcal{X}$) and lower case (e.g., $x$) respectively. All sets are denoted in calligraphic font (e.g., $\mathcal{X}$). We use superscripts to denote the vectors, like $X^n := (X_1, \ldots, X_n)$. All logarithms are base $e$. The set of all probability distributions on a finite set $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. Notation concerning the method of types follows [13]. Given a vector $x^n = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, the type or empirical distribution is denoted as $\overline{T}_{x^n}(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = a\}, a \in \mathcal{X}$. The set of types formed from length-$n$ sequences with alphabet $\mathcal{X}$ is denoted as $\mathcal{P}_n(\mathcal{X})$. Given $P \in \mathcal{P}_n(\mathcal{X})$, the set of all sequences of length $n$ with type $P$, the type class, is denoted as $\mathcal{T}_P^n$. We use $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{N}$ to denote the set of real numbers, non-negative real numbers, and natural numbers respectively. Given any number $a \in \mathbb{N}$, we use $[a]$ to denote the collection of natural numbers between 1 and $a$.

A. Preliminaries

Given a sequence of distributions $Q = (Q_1, \ldots, Q_M) \in \mathcal{P}(\mathcal{X})^M$, for each $i \in [M]$, define the following linear combination of KL Divergence terms between a single distribution and a mixture distribution

$$G_i(Q) := \sum_{j \in \mathcal{M}_i} D \left( Q_j \left\| \frac{\sum_{i \in \mathcal{M}_i} Q_i}{M - 1} \right. \right),$$

where $\mathcal{M}_i$ is defined as the set of integers in $[M]$ except $i$. The distance measure $G_i(Q)$ will be used to construct our test.

To present our results on the characterization of the tradeoff among the probabilities of misclassification error, false alarm and false reject, several definitions that generalize the divergence are needed. Given any pair of distributions $(P_N, P_A)$, for any $x \in \mathcal{X}$, define the two information densities (log likelihood ratios):

$$t_1(x|P_N, P_A) := \log \frac{(M - 1)P_A(x)}{(M - 2)P_N(x) + P_A(x)},$$

$$t_2(x|P_N, P_A) := \log \frac{(M - 1)P_N(x)}{(M - 2)P_N(x) + P_A(x)}.$$

The following linear combinations of the expectations and variances of these two information densities $p_i$ are critical in presenting our main results:

$$GD_M(P_N, P_A) := E_{P_A}[t_1(X|P_N, P_A)] + (M - 2)E_{P_N}[t_2(X|P_N, P_A)],$$

$$V_M(P_N, P_A) := \text{Var}_{P_A}[t_1(X|P_N, P_A)] + (M - 2)\text{Var}_{P_N}[t_2(X|P_N, P_A)].$$

Furthermore, we need the following covariance function of the information densities

$$\text{Cov}_M(P_N, P_A) := -(GD_M(P_N, P_A))^2 + E_{P_A} \left[ (t_1(X|P_N, P_A))^2 \right] + 2(M - 2)E_{P_N}[t_1(X|P_N, P_A)]E_{P_N}[t_2(X|P_N, P_A)]$$

$$+ (M^2 - 5M + 7)(E_{P_N}[t_2(X|P_N, P_A)])^2 + (M - 3)E_{P_N} \left[ (t_2(X|P_N, P_A))^2 \right].$$

Then, the covariance matrix $V_M(P_N, P_A) = \{V_{i,j}(P_N, P_A)\}_{(i,j) \in [M-1]^2}$ is defined as

$$V_{i,j}(P_N, P_A) = \begin{cases} V_M(P_N, P_A) & \text{if } i = j \\ \text{Cov}_M(P_N, P_A) & \text{otherwise}. \end{cases}$$
For any $k \in \mathbb{N}$, $Q_k(x_1, \ldots, x_k; \mu, \Sigma)$ is the multivariate generalization of the complementary Gaussian cdf defined as follows:

$$Q_k(x_1, \ldots, x_k; \mu, \Sigma) := \int_{x_1}^{\infty} \cdots \int_{x_k}^{\infty} \mathcal{N}(x; \mu; \Sigma)dx,$$

where $\mathcal{N}(x; \mu; \Sigma)$ is the pdf of a $k$-variate Gaussian with mean $\mu$ and covariance matrix $\Sigma$ [14]. Furthermore, for any $k \in \mathbb{N}$, we use $1_k$ to denote a row vector of length $k$ with all elements being one and we use $\Theta_k$ similarly. The complementary Gaussian cdf with covariance matrix $V_M(P_N, P_A)$ and mean values $G_{DM}(P_N, P_A)$ bounds the probability of false reject.

Finally, given any $\lambda \in \mathbb{R}_+$ and any pair of distributions $(P_N, P_A)$, for each $i \in [M]$, define the following quantity

$$LD_i(\lambda, P_N, P_A) := \min_{(j, k) \in [M]^2: j \neq k} \min_{G_j(Q) \leq \lambda, G_k(Q) \leq \lambda} \left( D(Q_i\|P_A) + \sum_{l \in \cal{M}_i} D(Q_l\|P_N) \right),$$

(9)

The above quantity is key to characterize the exponential decay rate of the false reject probability.

### B. Problem Formulation

We start by assuming that there is at most one outlying sequence. Consider a set of $M$ observed sequences $X^n := \{X^n_1, \ldots, X^n_M\}$ and a pair of nominal distribution $P_N$ and an anomalous distribution $P_A$ defined on the finite alphabet $\cal{X}$. All sequences, with at most one exception, are generated i.i.d. from $P_N$. The goal of outlying hypothesis testing is to discern the outlier that is generated i.i.d. from the anomalous distribution $P_A \in \mathcal{P}(\cal{X})$ if an outlier is present. Throughout this paper, we assume that both the nominal distribution $P_N$ and the anomalous distribution $P_A$ are unknown. Furthermore, to avoid degenerated cases, similar to [2], we consider only distributions $(P_N, P_A)$ with identical supports.

Under this setting, the objective of detecting the potential outlier is equivalent to making a correct decision in the $(M+1)$-ary hypothesis testing problem with following hypotheses:

- $H_i, i \in [M]$: the $i$-th sequence $X^n_i$ is the outlying sequence, i.e., $X^n_i \sim P_A$ and $X^n_j \sim P_N$ for all $j \in [M] \setminus \{i\}$;
- $H_r$: there is no outlying sequence, i.e., $X^n_i \sim P_N$ for all $j \in [M]$,

where $H_r$ denotes the null hypothesis.

The main task in the above OHT problem is to design a decision rule (test) $\phi_n : X^{Mn} \rightarrow \{H_1, \ldots, H_M, H_r\}$ with “good” performance be specified. Any test $\phi_n$ partitions the sample space $X^{Mn}$ into $M + 1$ disjoint regions: $\{A_i(\phi_n)\}_{i \in [M]}$ where $X^{Mn} \in A_i(\phi_n)$ favors the non-null hypothesis $H_i$ and a reject region $A_r(\phi_n) := \bigcup_{i \in [M]} A_i(\phi_n)^c$ where $X^{Mn} \in A_r(\phi_n)$ favor the null hypothesis $H_r$.

Given any test $\phi_n$ and any pair of nominal and anomalous distributions $(P_N, P_A) \in \mathcal{P}(\cal{X})^2$, the performance of the test $\phi_n$ is evaluated by the following misclassification error, false reject and false alarm probabilities:

$$\beta_i(\phi_n, P_N, P_A) := \mathbb{P}_r\{\phi_n(X^n) \notin H_i, H_r\}, \quad i \in [M],$$

(10)

$$\zeta_i(\phi_n, P_N, P_A) := \mathbb{P}_r\{\phi_n(X^n) = H_r, \quad i \in [M],$$

(11)

$$\delta_i(\phi_n, P_N, P_A) := \mathbb{P}_r\{\phi_n(X^n) \notin H_i\},$$

(12)

where for each $i \in [M]$, we define $\mathbb{P}_r\{\cdot\} := \mathbb{P}\{\cdot | H_r\}$ where $X^n_i$ is distributed i.i.d. according to $P_A$ and for $X^n_j$ is distributed according to $P_N$ for each $j \in [M] \setminus \{i\}$ and we define $\mathbb{P}_r\{\cdot\} := \mathbb{P}\{\cdot | H_i\}$ where all sequences are generated i.i.d. from $P_N$ for all $i \in [M]$. Consistent with the literature on hypothesis testing (e.g., [15]), we define $\beta_i(\phi_n, P_N, P_A)$ and $\zeta_i(\phi_n, P_N, P_A)$ as type-$i$ misclassification error and false reject probabilities, respectively, and we define $\delta_i(\phi_n, P_N, P_A)$ as the false alarm probability. Our main results characterized the tradeoff among the probabilities of misclassification error in (10), false rejection in (11) and false alarm in (12) under different settings.

### C. A Universal Test

Throughout the section, we use a threshold-based test that takes the types of each of the observed sequences as inputs and it outputs a decision among the $(M+1)$ hypotheses. Given $M$ observed sequences $x^n = (x^n_1, \ldots, x^n_M)$ and any positive real number $\lambda \in \mathbb{R}_+$, the universal test operates as follows:

$$\psi_n(x^n) := \begin{cases} H_2 & \text{if } S_i(x^n) < \min_{j \in [M]} S_j(x^n) \text{ and } \min_{j \in [M]} S_j(x^n) > \lambda \\ H_r & \text{otherwise}, \end{cases}$$

(13)

where the scoring function $S_i(x^n)$ is

$$S_i(x^n) := G_i(\hat{T}_x^n_1, \ldots, \hat{T}_x^n_M),$$

(14)

where $G_i(\cdot)$ is the function defined in (1) as the sum of the KL divergence of the empirical distribution of each sequence $x^n_j$, with the exception of $x^n_i$. Note that the threshold $\lambda$ can be replaced with a function of sequence length $n$, e.g., $\lambda_n$ to emphasize the dependence of the threshold on $n$. This way, all our results still hold but necessary limit operations are required.
We first informally explain the test in (14) from an asymptotic point of view. Intuitively, if \( x^n_j \) is the anomalous sequence that is generated from the unknown distribution \( P_A \), then as the length of each observed sequence \( n \) increases, using the weak law of large numbers, we know that the empirical distribution of \( x^n_j \) tends to \( P_A \) and the empirical distribution \( x^n_j \) for each \( j \in \mathcal{M} \), tends to the nominal distribution \( P_N \). Thus, the scoring function \( S_i(x^n) \) tends to zero and the scoring function of \( S_j(x^n) \) for each \( j \in \mathcal{M} \) tends to \( GD_M(P_N, P_A) \), which is strictly larger than zero if \( P_N \neq P_A \). Therefore, for any threshold \( \lambda \) that is positive but essentially less than \( GD_M(P_N, P_A) \), with high probability, it is possible to identify the outlier if it exists. On the other hand, if there is no outlier, then with the same logic, for each \( i \in [M] \), the scoring function \( S_i(x^n) \) tends to zero and naturally the null hypothesis is decided for any positive threshold \( \lambda \). Therefore, the test in (13) is consistent asymptotically for any \( P_N \neq P_A \) such that the threshold \( \lambda < GD_M(P_N, P_A) \).

Next we explain why the test is universal with guarantees on all three error probabilities. Note that the threshold \( \lambda \) can be arbitrary and it is independent of the unknown nominal and anomalous distributions \( (P_N, P_A) \). This is the reason why the above test is universal. However, the performance of the test under a particular pair of \( (P_N, P_A) \) does depend on \( (P_N, P_A) \).

Our main contribution in this section lies in the analysis of the performance of the universal test in (1) for any \( (P_N, P_A) \). Specifically, we show that both misclassification and false alarm probabilities decay exponentially fast, essentially with the speed of the threshold \( \lambda \), and we bound the false reject probabilities under two different regimes. Furthermore, we demonstrate the optimality of the above test for both regimes in a generalized Neyman-Pearson sense to be specified.

Finally, we remark that the scoring function \( S_i(x^n) \) was also used in [2, Eq. (15)] to construct a test for the case without the null hypothesis. However, for the case with the null hypothesis, a threshold-based test was given in [2, Eq. (25)] where the test relies on the pairwise difference of log likelihoods of the joint empirical distributions under different hypotheses. Note that the test in [2, Eq. (25)] is different from our test in (13) that uses different empirical distributions.

D. Second-Order Asymptotic Approximation to the Non-Asymptotic Performance

Our first set of results concern the characterization of the performance of the universal test in (13) in terms of probabilities of misclassification error, false alarm and false reject probabilities in the second-order asymptotic regime. We first demonstrate a non-asymptotic achievability result and then prove that such a result is optimal up to second-order, in a generalized Neyman-Pearson sense, in the spirit of [3], [4].

1) Achievability:

**Theorem 1.** For every pair of nominal and anomalous distributions \( (P_N, P_A) \in \mathcal{P}(X)^2 \), given any positive real number \( \lambda \in \mathbb{R}_+ \), the universal test in (13) satisfies

\[
\max_{i \in [M]} \beta_i(\psi_n | P_N, P_A) \leq \exp \left( -n \lambda + |X| \log((M-1)n + 1) \right),
\]

(15)

\[
P_{fa}(\psi_n | P_N, P_A) \leq M(M-1) \exp(-n \lambda + |X| \log((M-1)n + 1)),
\]

(16)

\[
\max_{i \in [M]} \zeta_i(\psi_n | P_N, P_A) \leq 1 - Q_{M-1} \left( \sqrt{n} \left( \lambda - GD_M(P_N, P_A) + O \left( \frac{\log n}{n} \right) \right) \right) \times 1_{M-1:0_{M-1}; V_M(P_N, P_A)} + O \left( \frac{1}{\sqrt{n}} \right),
\]

(17)

The proof of Theorem 1 is provided in Appendix A. We make several remarks.

For any finite number of observed sequences \( M \), as the length of each observed sequence \( n \) increases, both the maximal classification error (cf. (10)) and the false alarm (cf. (12)) probabilities decay exponentially fast with a speed lower bounded given by the threshold \( \lambda \) in the universal test in (13), i.e.,

\[
\lim_{n \to \infty} \frac{1}{n} \min_{i \in [M]} \left\{ \min \left\{ -\log \beta_i(\psi_n | P_N, P_A), -\log P_{fa}(\psi_n | P_N, P_A) \right\} \right\} \geq \lambda.
\]

(18)

Furthermore, asymptotically, the upper bound on the maximal false reject probability (cf. (11)) converges to

\[
\lim_{n \to \infty} \left( 1 - Q_{M-1} \left( \sqrt{n} \left( \lambda - GD_M(P_N, P_A) \right) \times 1_{M-1:0_{M-1}; V_M(P_N, P_A)} \right) \right),
\]

which is a function of the threshold \( \lambda \) and the pair of distributions \( (P_N, P_A) \). To better understand the seemingly complicated upper bound on the false reject probability, for any \( \varepsilon \in (0, 1) \), we can define

\[
L_M^*(\varepsilon | P_N, P_A) := \max \left\{ L \in \mathbb{R} : Q_{M-1} \left( L \times 1_{M-1:0_{M-1}; V_M(P_N, P_A)} \right) \geq 1 - \varepsilon \right\},
\]

(19)

Then the upper bound in (17) implies that if the threshold \( \lambda \) satisfies that

\[
\lambda \leq GD_M(P_N, P_A) + \frac{L_M^*(\varepsilon | P_N, P_A)}{\sqrt{n}} =: \lambda^*(n, \varepsilon | P_N, P_A),
\]

(20)

then the maximal false reject probability under \( (P_N, P_A) \) is asymptotically upper bounded by \( \varepsilon \), for any \( \varepsilon \in (0, 1) \), say as small as \( 10^{-10} \). If the threshold \( \lambda \) further satisfies that \( \lambda < GD_M(P_N, P_A) \), then the false reject probability converges to zero. As we shall show later (in Theorem 3), actually, if \( \lambda < GD_M(P_N, P_A) \), the false reject probability converges to zero exponentially fast with a speed lower bounded by an explicit function of the threshold \( \lambda \).
Fig. 1. Illustration of phase transition for our test. Here we consider Bernoulli sources $P_N = \text{Bern}(0.2)$ and $P_A = \text{Bern}(0.4)$. We assume that there are $M = 4$ observed sequences and one sequence is the outlier. We plot the maximal false reject probability, i.e., $\max_{i \in [n]} \zeta_i(\psi_n | P_N, P_A)$.

Note that $\lambda^*(n, \varepsilon | P_N, P_A)$ is a critical bound for the threshold in the test, which trades off a lower bound $\lambda$ on the exponential decay rates of misclassification error and false alarm probabilities and a non-vanishing upper bound $\varepsilon \in (0, 1)$ for the maximal false reject probability. Such a result is known as a second-order asymptotic result in information theory since it provides the second dominant term $L_M^*(\varepsilon | P_N, P_A)$ beyond the leading constant term $\text{GD}_M(P_N, P_A)$ asymptotically as $n \to \infty$. Furthermore, as shown in non-asymptotic analysis for channel coding [16], second-order asymptotic results often provide good approximation to the performance for finite length $n$. We provide a numerical example to illustrate the validity of this claim in Section II-D3. Furthermore, the result in (20) implies a phase transition phenomenon for our universal test. In particular, if the threshold $\lambda$ is greater than $\text{GD}_M(P_N, P_A)$, then asymptotically the false reject probabilities tend to one. On the other hand, if $\lambda < \text{GD}_M(P_N, P_A)$, then asymptotically the false reject probabilities vanish. See Figure 1 for a numerical illustration.

Theorem 1 also captures the influence of the number of sequences $M$ on the performance of the universal test (13). To study the asymptotic case of $M \to \infty$, we need to make an assumption on the order of $M$ and $n$. In fact, as long as $\lim_{n \to \infty} \frac{\log(M)}{n} \to 0$, the asymptotic lower bound in (18) holds. Intuitively, when one has a larger number of sequences, it should be easier to learn the nominal distribution and thus achieve better performance. This should imply that as $M$ increases, the upper bound $\lambda^*(n, \varepsilon | P_N, P_A)$ on the homogeneous error exponent $\lambda$ in (20) increases as well. To verify this intuition, the second-order result in (20) for Bernoulli distributions with different values of $M$ is plotted in Figure 2. The influence of $M$
on the performance of the universal test in (13) is dominated by $\text{GD}_M(P_N, P_A)$. In fact,

$$\frac{\partial \text{GD}_M(P_N, P_A)}{\partial M} = D \left( P_N \left\| \frac{(M-2)P_N + P_A}{M-1} \right\| \right) > 0. \quad (21)$$

Thus, as the number of sequences $M$ increases, the performance of the test in (13) improves. In the extreme case, as $M \to \infty$, we have

$$\lim_{M \to \infty} \text{GD}_M(P_N, P_A) = D(P_A \| P_N). \quad (22)$$

This implies that the maximum asymptotic decay rate of the misclassification and false alarm probabilities of the universal test under any pair of nominal and anomalous distributions $(P_N, P_A)$ in (13) is $D(P_A \| P_N)$ as the number of sequences $M$ tends to infinity, assuming that the false reject probability does not tend to one.

Finally, we remark that Theorem 1 is relevant to $M$-ary hypothesis testing with empirically observed statistics [3], [4], [11]. In fact, our proof technique for Theorem 1 can be used to strengthen [4, Theorem 4.1] by removing the condition in [4, Section 4.2] on the uniqueness of the minimizing distribution for the scoring function in [4, Eq. (4.4)]. However, the results in [3], [4], [11] do not directly apply to our setting since the two problems are significantly different. In the $M$-ary hypothesis testing problem with empirically observed statistics, one is given $M$ training sequences and one test sequence. The task there is to identify the true distribution of the test sequence among the empirical distributions of the training sequences. In contrast, in the outlying sequence detection problem addressed in Theorem 1, we are given $M$ sequences and our task is to identify the potential outlier if it exists.

2) Converse: With the above achievability result on the performance of the universal test in (13), it remains to show that the test is in fact optimal in the generalized Neyman-Pearson sense. To prove the following, we use arguments similar to those of Gutman [3].

**Theorem 2.** Given any positive real number $\lambda \in \mathbb{R}_+$, let the test $\phi_n$ satisfy the following for all pairs of nominal and anomalous distributions $(\hat{P}_N, \hat{P}_A)$,

$$\max \{ \max_{i \in \{M\}} \beta_i(\phi_n | \hat{P}_N, \hat{P}_A), P_{\text{fa}}(\phi_n | \hat{P}_N, \hat{P}_A) \} \leq \exp(-n\lambda). \quad (23)$$

Then for any pair of nominal and anomalous distributions $(P_N, P_A)$, the minimal false reject probability satisfies

$$\min_{i \in \{M\}} \zeta_i(\phi_n | P_N, P_A) \geq 1 - Q_{M-1} \left( \sqrt{n} (\lambda - \text{GD}_M(P_N, P_A) + O \left( \frac{\log n}{n} \right)) \times 1_{M-1}; 0_{M-1}; V_M(P_N, P_A) \right) + O \left( \frac{1}{\sqrt{n}} \right). \quad (24)$$

The proof of Theorem 2 is provided in Appendix B.

The result in Theorem 2 holds for any number of observed sequences $M$ and when the length $n$ of each observed sequence $n$ is such that $O \left( \frac{\log n}{n} \right)$ and $O \left( \frac{1}{\sqrt{n}} \right)$ can be neglected. Furthermore, Theorem 2 implies that the test in (13) is optimal in the generalized Neyman-Pearson sense. Specifically, among all tests that ensure exponential decay of the maximal misclassification error and false alarm probabilities at a speed no less than $\lambda$, the universal test in (13) achieves the minimal false reject probability in a second-order asymptotic sense such that $\lim_{n \to \infty} Q_{M-1} \left( \sqrt{n} (\lambda - \text{GD}_M(P_N, P_A) + O \left( \frac{\log n}{n} \right)) \times 1_{M-1}; 0_{M-1}; V_M(P_N, P_A) \right) > 0$.

3) A Numerical Example: We present a numerical example to illustrate Theorem 1. Consider the binary alphabet $\mathcal{X} = \{0, 1\}$ and $M = 4$. Assume that there is exactly one outlier sequence and let $\text{Bern}(p)$ denote a Bernoulli distribution with parameter $p \in (0, 1)$. For any $(p, q) \in (0, 1)^2$ such that $p \neq q$, we set the nominal distribution $P_N$ as $\text{Bern}(p)$ and the anomalous distribution $P_A$ as $\text{Bern}(q)$. We make the above nominal and anomalous distribution assumptions in order to demonstrate tightness of the inequality (20) in the theorem. For the above example, the information densities (cf. (2) and (3)) satisfy

$$v_1(x | P_N, P_A) := \mathbb{I}(x = 0) \log \frac{(M-1)(1-q)}{(M-2)(1-p) + 1 - q} + \mathbb{I}(x = 1) \log \frac{(M-1)q}{(M-2)p + q}, \quad (25)$$

$$v_2(x | P_N, P_A) := \mathbb{I}(x = 0) \log \frac{(M-1)(1-p)}{(M-2)(1-p) + 1 - q} + \mathbb{I}(x = 1) \log \frac{(M-1)p}{(M-2)p + q}. \quad (26)$$

Furthermore,

$$\text{GD}_M(P_N, P_A) = D_b \left( q \left\| \frac{(M-2)p + q}{M-1} \right\| \right) + (M-2)D_b \left( p \left\| \frac{(M-2)p + q}{M-1} \right\| \right), \quad (27)$$

where $D_b(p|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is the binary KL divergence function. The variance $V_M(P_N, P_A)$ is given by

$$V_M(P_N, P_A) = E_{P_N} [(v_1(X | P_N, P_A))^2] + (M-2)E_{P_N} [(v_2(X | P_N, P_A))^2] - \left( D_b \left( q \left\| \frac{(M-2)p + q}{M-1} \right\| \right)^2 - (M-2) \left( D_b \left( p \left\| \frac{(M-2)p + q}{M-1} \right\| \right)^2 \right). \quad (28)$$
To verify the tightness of our theoretical result, we simulate the false reject probability under the nominal distribution is tight for this numerical example. For this purpose, the nominal distribution is assumed to be \( P_N = \text{Bern}(0.2) \) and the anomalous distribution is assumed to be \( P_A = \text{Bern}(0.4) \). The error bar denotes two standard deviations below and above the mean value. As observed, the simulated false reject probability approaches the target value \( \varepsilon = 0.1 \) as the lengths of observed sequences become moderate. This implies that our theoretical characterization in Theorem 1 is tight for this numerical example.

Similarly, we can also calculate \( \text{Cov}_M(P_N, P_A) \) (cf. (6)) and thus the covariance matrix \( V_M(P_N, P_A) \).

For the case of \( p = 0.2, q = 0.4 \) and \( M = 4 \), we have

\[
V_M(P_N, P_A) = \begin{bmatrix}
0.1331 & 0.1106 & 0.1106 \\
0.1106 & 0.1331 & 0.1106 \\
0.1106 & 0.1106 & 0.1331
\end{bmatrix}.
\] (29)

In Figure 3, the simulated false reject probability of our test in (13) is plotted versus a target value \( \varepsilon = 0.1 \) for the case where there is only one outlying sequence out of \( M = 4 \) observed sequences. For each \( n \in \{100, 125, \ldots, 200, 300, \ldots, 1500\} \), the threshold in the universal test in (13) is chosen as

\[
\lambda = \text{GD}_M(P_N, P_A) + \frac{L^*_M(\varepsilon|P_N, P_A)}{\sqrt{n}}.
\] (30)

From the theoretical result in Theorem 1 and the implication in (20), such a choice of the threshold ensures that the false reject probability is upper bounded by the parameter \( \varepsilon \) for any pair of anomalous distributions \( (\hat{P}_N, \hat{P}_A) \) such that \( \lambda < \text{GD}_M(\hat{P}_N, \hat{P}_A) + \frac{L^*_M(\varepsilon|P_N, P_A)}{\sqrt{n}} \). In particular, when \( (\hat{P}_N, \hat{P}_A) = (P_N, P_A) \), the false reject probability should be roughly \( \varepsilon \). To verify the tightness of our theoretical result, we simulate the false reject probability under \( (P_N, P_A) \). In our simulation, for each data point, our test is run independently \( 10^7 \) times and the false reject probability is calculated as the ratios of false alarms. From Figure 3, we observe that the simulated false reject probability meets the desired value when the length of the sequences is moderate. This illustrates that the second-order asymptotic result in (20) provides a good approximation to the performance of the universal test for moderate \( n \).

E. Asymptotic Decay Rates

For accurate anomaly detection, all error probabilities should be small to ensure that no outlier is missed or identified incorrectly. Thus, a constant or even vanishing false reject probability might not suffice when the length of the observed sequence \( n \) is extremely large. In the following theorem, we obtain an asymptotic tradeoff between the exponents of false reject probabilities and the homogeneous error exponent for misclassification error and false alarm probabilities. Recall the definition of \( \text{LD}_i(\lambda, P_N, P_A) \) in (9).

**Theorem 3.** For every pair of nominal and anomalous distributions \( (P_N, P_A) \), given any positive real number \( \lambda \in \mathbb{R}_+ \), the universal test in (13) satisfies:

\[
\lim_{n \to \infty} \frac{1}{n} \min_{i \in [M]} \left\{ \min_{\psi \in [M]} \{- \log \beta_i(\psi|P_N, P_A)\}, - \log P_{fa}(\psi|P_N, P_A) \right\} \geq \lambda,
\] (31)

\[
\lim_{n \to \infty} - \frac{1}{n} \log \zeta_i(\psi|P_N, P_A) \geq \text{LD}_i(\lambda, P_N, P_A).
\] (32)
Conversely, given any positive real number \( \lambda \in \mathbb{R}_+ \), for any test \( \phi_n \) such that for all pairs of nominal distributions \( (\hat{P}_N, \hat{P}_A) \),

\[
\lim_{n \to \infty} \frac{1}{n} \min \left\{ \min_{i \in [M]} \{- \log \beta_i(\phi_n | \hat{P}_N, \hat{P}_A)\}, - \log P_{\text{fn}}(\psi_{\text{a}} | \hat{P}_N, \hat{P}_A) \right\} \geq \lambda,
\]

(33)

under any pair of nominal and anomalous distributions \( (P_N, P_A) \), the false reject exponent satisfies

\[
\lim_{n \to \infty} -\frac{1}{n} \log \zeta_i(\phi_n | P_N, P_A) \leq \text{LD}_i(\lambda, P_N, P_A).
\]

(34)

The differences between the proof of Theorem 3 and the proofs of Theorems 1, 2 lie in the analysis of the false reject probability. See Appendix C.

To ensure that all three kinds of error probabilities decay exponentially, we need \( \min\{\lambda, \min_{i \in [M]} \text{LD}_i(\lambda, P_N, P_A)\} > 0 \). Given any \( (P_N, P_A) \), for each \( i \in [M] \), \( \text{LD}_i(\lambda, P_N, P_A) \) (cf. (9)) is non-increasing in \( \lambda \) and \( \text{LD}_i(\lambda, P_N, P_A) = 0 \) if and only if \( \lambda \geq \text{GD}_M(\hat{P}_N, \hat{P}_A) \) (cf. Appendix H for justification). Therefore, for any pair of nominal and anomalous distributions \( (P_N, P_A) \) such that \( \text{GD}_M(P_N, P_A) > 0 \) and \( \lambda < \text{GD}_M(P_N, P_A) \), all three kinds of error probabilities decay to zero exponentially fast. However, in practice, one cannot know either \( P_N \) or \( P_A \). Thus, the above result implies that one can choose a smaller threshold \( \lambda \) to ensure exponentially consistent performance under a larger set of distributions. When one has some information about the underlying true pair of distributions \( (P_N, P_A) \), one can choose a large enough \( \lambda \) to ensure good homogeneous error exponent and a positive false reject exponent.

We further discuss the tradeoff between the false reject exponent \( \text{LD}_i(\lambda, P_N, P_A) \) and the homogeneous error exponent \( \lambda \) under each hypothesis. Specifically, one might wonder what value is taken on by the largest false reject exponent given any positive \( \lambda \). In Appendix H, we show that

\[
\sup_{\lambda \in \mathbb{R}_+} \text{LD}_i(\lambda, P_N, P_A) < \min_{Q \in \mathcal{P}(\mathcal{X})} D(Q \parallel P_A) + (M - 1)D(Q \parallel P_N),
\]

(35)

and thus provide an answer to the above question. Note that the right hand side in (35) is positive if \( P_N \neq P_A \).

The converse part states that the universal test in (13) is also optimal in the generalized Neyman-Pearson sense when the false reject probabilities decay exponentially fast. Specifically, among all tests that ensure exponential decay of misclassification error and false alarm probabilities for all possible pairs of nominal and anomalous distributions, the universal test in (13) has the largest false reject exponent under any pair of nominal and anomalous distributions.

III. CASE OF MULTIPLE OUTLYING SEQUENCES

In this section, we generalize the results in Section II to the case of multiple outlying sequences where each outlying sequence can be generated according to a potentially different anomalous distribution. We assume that the number of outliers is unknown but less than half of total number of the observed sequences and then study the performance a universal test, which generalizes (13). Furthermore, we demonstrate the optimality of such test in the generalized Neyman-Pearson sense.

A. Preliminaries

We first present necessary definitions. Note that the definitions here generalize those in Section II-A.

Given any \( M \in \mathbb{N} \), let \( [T] := \frac{M}{2} - 1 \). For any \( t \in [T] \), let \( S \) denote the set of all subsets of \( [M] \) whose cardinality (size) is \( t \), i.e.,

\[
S := \{ B \subseteq [M] : |B| = t \}.
\]

(36)

Then, define the union of subsets \( S \) over \( t \in [T] \) as follows:

\[
S := \bigcup_{t \in [T]} S.
\]

(37)

Given a sequence of distributions \( Q = (Q_1, \ldots, Q_M) \in \mathcal{P}(\mathcal{X})^M \) and each \( B \in S \), analogously to (1), define the following linear combination of KL divergence terms

\[
G_B(Q) := \sum_{i \in \mathcal{M}_B} D \left( Q_i \parallel \frac{\sum_{i \in \mathcal{M}_B} Q_i}{M - |B|} \right),
\]

(38)

where \( \mathcal{M}_B \) is defined as the set of elements that are in \( [M] \) but not in \( B \), i.e., \( \mathcal{M}_B := [M] \setminus B = \{ i \in [M] : i \notin B \} \).

For any \( i \in B \), define \( J_B(i) := j \) if \( i \) is the \( j \)-th largest element in \( B \). Given any \( B \in S \) and any tuple of distributions \( P_B = (P_N, P_{A,1}, \ldots, P_{A,|B|}) \in \mathcal{P}(\mathcal{X})^{|B|+1} \), for any two sets \( (B, C) \in S^2 \), define the following mixture distribution

\[
P_{\text{Mix}}^{(B,C,P_N,P_E)}(x) := \frac{1}{M - |C|} \left( \sum_{i \in (B \setminus \mathcal{M}_C)} P_{A,J_B(i)}(x) + \sum_{i \in (\mathcal{M}_B \setminus \mathcal{M}_C)} P_N(x) \right),
\]

(39)
and, parallel to (2) and (3), define the following information densities (log likelihoods)

\[ t_{1,l}(x|B, C, P_N, P_B) := \log \frac{P_{A,l}(x)}{P_{Mix}^{(B,C,P_N,P_B)}(x)}, \quad l \in [|B|], \]  

\[ t_2(x|B, C, P_N, P_B) := \log \frac{P_N(x)}{P_{Mix}^{(B,C,P_N,P_B)}(x)}. \]  

Analogously to (7), define the covariance matrix \( Q \)

\[ \text{Cov}(P_A(j), P_B(k)) := \mathbb{E}_{P_{A,j}(i)}[t_{1,j}(i) t_{1,j}(i)] + \sum_{i \in (M_B \setminus M_C)} \mathbb{E}_{P_{N}}[t_2(X|B, C, P_N, P_B)] \]

Analogously to (4) to (6), define the following linear combinations of expectations and variances of information densities:

\[ GD(B, C, P_N, P_B) := \sum_{i \in (B \setminus M_C)} \mathbb{E}_{P_{A_j}(i)}[t_{1,j}(i)(X|B, C, P_N, P_B)] + \sum_{i \in (M_B \setminus M_C)} \mathbb{E}_{P_N}(t_2(X|B, C, P_N, P_B)) \]

\[ V(B, C, P_N, P_B) := \sum_{i \in (B \setminus M_C)} \text{Var}_{P_{A_j}(i)}[t_{1,j}(i)(X|B, C, P_N, P_B)] + \sum_{i \in (M_B \setminus M_C)} \text{Var}_{P_N}(t_2(X|B, C, P_N, P_B)). \]  

For simplicity, given any \((B, C) \in S^2\) and any variables \((x_1, \ldots, x_M)\), let

\[ t_{B,C}(x_1, \ldots, x_M|P_N, P_B) := \sum_{j \in (B \setminus M_C)} t_{1,j}(j)(x_j|B, C, P_N, P_B) + \sum_{j \in (M_B \setminus M_C)} t_2(x_j|B, C|P_N, P_B). \]  

For ease of latter presentation, let \( S_B \) denote the set \( S \setminus \{B\} \), i.e., \( \{C \in S : C \neq B\} \). Furthermore, let the elements in \( S_B \) be ordered as \( \{C_1, \ldots, C_{|S| - 1}\} \). Then for each \((i, k) \in [|S| - 1]^2\) such that \( i \neq k \), define the covariance

\[ \text{Cov}(C_i, C_k, P_N, P_B) := \mathbb{E}_{P_{B,C}(i)}(X_1, \ldots, X_M|P_N, P_B) t_{B,C}(X_1, \ldots, X_M|P_N, P_B) \]

Analogously to (8), define a covariance matrix \( V(B, P_N, P_B) = \{V_{i,k}(B, P_N, P_B)\}_{(i,k) \in [|S| - 1]^2} \) where

\[ V_{i,k}(B, P_N, P_B) = \begin{cases} V(B, C_i, P_N, P_B), & \text{if } i = k, \\ \text{Cov}(C_i, C_k, P_N, P_B), & \text{otherwise}. \end{cases} \]  

The complementary cdf \( Q_k(-) \) in (8), together with \( GD(B, C, P_N, P_B) \) and \( V(B, P_N, P_B) \), will be critical to upper bound the false reject probabilities.

Finally, given any \( \lambda \in \mathbb{R}_+ \) and any tuple of distributions \( P_B = (P_N, P_{A,1}, \ldots, P_{A,T}) \in P_T(\mathcal{X}) \), for each \( B \in S \), define the following quantity:

\[ \text{LD}_B(\lambda, P_N, P_B) := \min_{(C, D) \in S^2 : C \neq D} \min_{Q \in (P(\mathcal{X}))^{M_B}} \min_{G_C(Q) \leq \lambda, G_D(Q) \leq \lambda} \left( \sum_{i \in B} D(Q_i \parallel P_{A,i}(i)) + \sum_{i \in M_B} D(Q_i \parallel P_N) \right). \]  

The quantity \( \text{LD}_B(\lambda, P_N, P_B) \) will characterize the false reject exponent under each hypothesis.

### B. Problem Formulation

Assume that there are at most \( T = [\frac{M}{2} - 1] \) outlying sequences out of \( M \) observed sequences \( X = (X_1^n, \ldots, X_M^n) \). In the outlier hypothesis testing problem with at most \( T \) outliers, the task is to decide whether there are outliers and identify the set of outlying sequences if any exist. We assume that each outlying sequence is generated i.i.d. according to a possibly different anomalous distribution. Specifically, let \( P_T := (P_{A,1}, \ldots, P_{A,T}) \) be a collection of \( T \) anomalous distributions that are different from the nominal distribution \( P_N \), all defined on the finite alphabet \( \mathcal{X} \) with the same support. Furthermore, for any \( B \in S \), let \( P_B \) denote the collection of distributions \( (P_1, \ldots, P_{B}) \). When \( B \in S \) denotes the index of the outlying sequences, for any \( l \in B, X_l^n \) is generated i.i.d. from \( P_{A, j}(l) \). Recall that \( j_B(l) \) denotes \( l \)-th largest element in \( B \).

Recall that \( S = \bigcup_{l \in [T]} S_l \). Since the exact number of outliers is unknown, there are in total \( |S| + 1 = \sum_{l \in [T]} \binom{M}{l} + 1 \) possible configurations of outlying sequences. Formally, the task is to design a test \( \phi_n : \mathcal{X}^{Mn} \rightarrow \{\{H_B\}_{B \in S}, H_r\} \) to classify between the following \( |S| + 1 \) hypotheses:

- \( H_B \): where \( B \in S \): the set of outlying sequences are sequences \( X_j^n \) with \( j \in B \);
- \( H_r \): there is no outlying sequence.

Similarly to Section II, the null hypothesis is introduced to model the case when there is no outlier among all \( M \) observed sequences.
Given any test \( \phi_n \), under any tuple of nominal and anomalous distributions \((P_N, P_T) = (P_N, P_{A,1}, \ldots, P_{A,T})\), the performance of \( \phi_n \) is evaluated by the following misclassification error, false reject and false alarm probabilities:

\[
\begin{align*}
\beta_B(\phi_n|P_N, P_T) &:= P_B\{\phi_n(X^n) \notin \{H_B, H_T\}\}, \\
\zeta_B(\phi_n|P_N, P_T) &:= P_B\{\phi_n(X^n) = H_T\}, \\
P_{fa}(\phi_n|P_N, P_T) &:= P_{\bar{I}}\{\phi_n(X^n) \neq H_I\},
\end{align*}
\]

where \( B \in S \) denotes the set of indices of outliers, and we define \( P_B(\cdot) := \Pr\{\cdot|H_B\} \) where for \( i \in M_S \), \( X_i^n \) is generated i.i.d. from the nominal distribution \( P_N \) and for \( i \in B \), \( X_i^n \) is generated i.i.d. from an anomalous distribution \( P_{A,B(i)} \), finally we define \( P_{\bar{I}}(\cdot) := \Pr\{\cdot|H_I\} \), where all sequences are generated i.i.d. from the nominal distribution \( P_N \).

### C. A Universal Test

Throughout the section, we use a threshold-based universal test that takes the empirical distribution of each observed sequence as the input and outputs a decision among all hypotheses.

Given \( M \) observed sequences \( x^n = (x_1^n, \ldots, x_M^n) \) and any positive real number \( \lambda \), the universal test operates as follows:

\[
\begin{align*}
\Psi_n(x^n) := \begin{cases} 
H_B & \text{if } S_B(x^n) \leq \lambda < \min_{C \in S_B} S_C(x^n) \\
H_T & \text{otherwise},
\end{cases}
\end{align*}
\]

which measures the sum of KL divergence between the empirical distribution of each sequence \( x_i^n \) with \( i \notin B \) relative to the average of the empirical distribution of all sequences \( x_j^n \) where \( j \notin B \). For the special case of \( T = 1 \), the test in (51) reduces to the test in (13).

We first provide intuition into why the above test should be asymptotically consistent. Assume that \( B \in S \) denotes the set of indices of the outliers and for each \( l \in B \), the outlier \( x_i^n \) is generated i.i.d. from \( P_{A,B(l)} \). Asymptotically as \( n \) tends to infinity, for any sequence \( x_i^n \) where \( i \notin B \), the empirical distribution \( \hat{T}_x^n \) tends to the nominal distribution \( P_N \) and for each \( x_i^n \) where \( i \in B \), the empirical distribution \( \hat{T}_x^n \) tends to \( P_{A,B(i)} \). Therefore, given any \( C \in S_B \), the scoring function \( S_C(x^n) \) converges to \( GD(B, C, P_N, P_B) \) and \( S_B(x^n) \) converges to zero. Note that \( GD(B, C, P_N, P_B) > 0 \) for any \( P_B \) where \( P_{A,j} \neq P_N \) for all \( j \in |B| \). Thus, asymptotically, the set of outliers \( B \) can be identified if \( \lambda < \min_{C \in S_B} GD(B, C, P_N, P_B) \). On the other hand, when there is no outlier, for each \( B \in S \), the scoring function \( S_B(x^n) \) tends to zero and thus, with any positive threshold \( \lambda \), the null hypothesis is decided. Therefore, the universal test in (51) is consistent asymptotically for any set of outlier indices \( B \in S \) and for any tuple of distributions \( P_B \) where \( P_{A,j} \neq P_N \) for all \( j \in |B| \) such that the threshold \( \lambda < \min_{C \in S_B} GD(B, C, P_N, P_B) \).

We next explain briefly the universality of the test and its associated theoretical guarantees on the error probabilities. As in the previous case of at most one anomalous sequence, the test in (51) is universal since \( \lambda \) can be selected without knowledge of the distributions. However, the performance of such a test under a particular tuple of nominal and anomalous distributions does depend on \( P_B \) where \( B \in S \) denotes the set of indices of the outliers. Our contribution in this section is two fold. We first show that the universal test in (51) has misclassification error and false alarm probabilities that decay exponentially, with a lower bound on the decay rate depending on \( \lambda \). Furthermore, we bound the false reject probability as a function of \( \lambda \) in two regimes where the false reject probability is either upper bounded by a constant or decays exponentially. We also demonstrate the optimality of the test in (51) in the generalized Neyman-Pearson sense under both regimes.

The statistic in (38) was also used in [2, Eq. (37)] to construct a test when the number \( t \) of outliers is known and when there is no null hypothesis. In contrast, the universal test in (51) does not assume any knowledge of the number of outliers, and in addition, incorporates a null hypothesis to include the possibility of no outliers.

### D. Second-Order Asymptotic Approximation to the Non-Asymptotic Performance

Our first set of results concern the characterization of the performance tradeoff among misclassification error, false alarm and false reject probabilities. Specifically, we first provide an achievability result, where the performance of the universal test in (51) is characterized in terms of misclassification and false alarm probabilities that decay exponentially fast when the false reject probability is upper bounded by a function of the threshold \( \lambda \). Furthermore, we demonstrate the optimality of the test in (51) in the generalized Neyman-Pearson sense.
Theorem 4. For any nominal distribution $P_N$ and anomalous distributions $P_T = (P_{A,1}, \ldots, P_{A,T})$, given any positive real number $\lambda \in \mathbb{R}_+$, the universal test in (51) satisfies that for each $B \in S$,

$$\beta_B(\Psi_n | P_N, P_T) \leq \exp \left( -n\lambda + |X| \log((M - 1)n + 1) \right),$$  \hspace{1cm} (53)

$$P_{ta}(\Psi_n | P_N, P_T) \leq |S|^2 \exp \left( -n\lambda + |X| \log((M - 1)n + 1) \right),$$  \hspace{1cm} (54)

$$\zeta_B(\Psi_n | P_N, P_T) \leq 1 - Q_{|S|-1}\left( \sqrt{n}\hat{\mu}(\lambda, P, P_B) ; 0_{|S|-1}, 0_{|S|-1}, V(B, P, P_B) \right) + O\left( \frac{1}{\sqrt{n}} \right),$$  \hspace{1cm} (55)

where $\hat{\mu}(\lambda, P, P_B)$ denotes the vector $(\lambda - \text{GD}(B, C_1, P_N, P_B) + O(\log n/n), \ldots, \lambda - \text{GD}(B, C_{|S|-1}, P_N, P_B) + O(\log n/n))$.

The proof of Theorem 4 is a generalization of the proof of Theorem 1 and is given in Appendix D.

Similarly to the result given in Theorem 1, when the number of outliers $M$ is finite, both misclassification and false alarm probabilities decay exponentially fast, with a speed lower bounded by $\lambda$ asymptotically when $n$ tends to infinity. On the other hand, the false reject under each hypothesis $H_B$ is upper bounded by a function of $\lambda$ and critical quantities $\text{GD}(B, C, P_N, P_B)$ and $V(B, P_N, P_B)$. Note that the threshold $\lambda$ trades off the lower bound on the decay rate of the homogeneous error exponent of the misclassification and false alarm probabilities and the upper bound on the false reject probability. If $\lambda$ increases, the homogeneous error exponent increases while the false reject probability increases as well. This implies that better performance in misclassification error and false alarm probabilities leads to a worse in false reject probabilities.

Asymptotically as $n \to \infty$, if the threshold $\lambda < \min_{i \in [|S|-1]} \text{GD}(B, C_i, P_N, P_B)$, then the false reject probability under hypothesis $H_B$ vanishes. One might also be interested in the more practical non-asymptotic case where $n$ is finite. Obtaining the exact solution to such case is almost impossible. However, a second-order asymptotic approximation to the non-asymptotic performance is possible using the result in (55). For this purpose, we define

$$\text{GD}(B, P_N, P_B) := \min_{i \in [|S|-1]} \text{GD}(B, C_i, P_N, P_B),$$  \hspace{1cm} (56)

as the minimum value of the vector $(\text{GD}(B, C_1, P_N, P_B), \ldots, \text{GD}(B, C_{|S|-1}, P_N, P_B))$ and let $d(B)$ be the number of elements in the vector that equals the minimal value, i.e., $d(B) := |\{ i \in [|S|-1] : \text{GD}(B, C_i, P_N, P_B) = \text{GD}(B, P_N, P_B) \}|$. Analogously to (19), given any $\varepsilon \in (0, 1)$, let

$$L^*(\varepsilon | B, P_N, P_B) := \max \left\{ L \in \mathbb{R} : Q_{d(B)}(L \times 1_{d(B)} ; 0_{d(B)}) ; V(B, P_N, P_B) \right\} \geq 1 - \varepsilon \bigg\}.$$  \hspace{1cm} (57)

If the threshold $\lambda$ satisfies that

$$\lambda \leq \text{GD}(B, P_N, P_B) + \frac{L^*(\varepsilon | B, P_N, P_B)}{\sqrt{n}},$$  \hspace{1cm} (58)

then as the blocklength $n$ increases, the upper bound on the false reject probability tends to $\varepsilon \in (0, 1)$. The second-order asymptotic upper bound in (58) provides further characterization beyond the first-order asymptotic constant term $\text{GD}(B, P_N, P_B)$. The second-order asymptotic upper bound on $\lambda$ trades off the homogeneous error exponent with any non-vanishing false reject probability $\varepsilon \in (0, 1)$ beyond the vanishing case with $\varepsilon \to 0$ implied by a first-order asymptotic analysis.

Finally, we discuss the influence of the number of observed sequences $M$ on the performance of the universal test in (51). As demonstrated in the above remark, $\text{GD}(B, P_N, P_B)$ is the critical quantity that is related with the performance of the test. Thus, it suffices to study the properties of $\text{GD}(B, P_N, P_B)$ as a function of $M$ under each hypothesis $H_B$. However, it is challenging to obtain closed form equations for the dependence of $\text{GD}(B, P_N, P_B)$ on $M$ when each outlier is generated from a unique anomalous distributions. Thus, we specialize our results to the case where all anomalous distributions are the same and given by $P_A$. Under this assumption, one can verify that

$$\text{GD}(B, P_N, P_B) = \min_{t \in [T]} \min_{l \in [B]} \left( |ID(P_A || P_{t,l,\text{Mix}}^t) + (M - t - l)D(P_N || P_{t,l,\text{Mix}}^t) \right),$$  \hspace{1cm} (59)

where $P_{t,l,\text{Mix}}^t = \frac{tP_A + (M - t - l)P_N}{M - t}$. For any $(t, l) \in [T] \times [B]$, one can verify that

$$\frac{\partial \text{GD}(B, P_N, P_B)}{\partial M} = \frac{1}{2}D(P_N || P_{t,l,\text{Mix}}^t).$$  \hspace{1cm} (60)

Thus, $\text{GD}(B, P_N, P_B)$ increases in $M$ if $D(P_N || P_{t,l,\text{Mix}}^t) > 0$, which holds almost surely for all distinct nominal and anomalous distributions. This implies that the performance of the universal test in (51) increases as the number of observed sequences $M$ increases when the number of outliers $|B|$ remains unchanged. On the other hand, the result in (59) implies that for a fixed number of observed sequences $M$, the performance of the test in (51) degrades as the number of outliers $|B|$ increases.

In the following theorem, it is shown that the universal test in (51) is optimal in the generalized Neyman-Pearson sense for second-order asymptotic analysis.
Theorem 5. Given any $\lambda \in \mathbb{R}_+$, for any test $\phi_n$ such that for all tuples of nominal and anomalous distributions $(\tilde{P}_N, \tilde{P}_T)$,

$$\beta_B(\phi_n|\tilde{P}_N, \tilde{P}_T) \leq \exp(-n\lambda),$$

(61)

then for any tuple of nominal and anomalous distributions $(P_N, P_T)$, for each $B \in \mathcal{S}$,

$$\zeta_B(\Psi_n|P_N, P_T) \geq 1 - Q_{|\mathcal{S}|-1}(\sqrt{n}\tilde{\mu}(\lambda, P_N, P_B); 0_{|\mathcal{S}|-1}; 0_{|\mathcal{S}|-1}; V(B, P_N, P_B)) + O\left(\frac{1}{\sqrt{n}}\right).$$

(62)

The proof of Theorem 5 is similar to that of Theorem 2 and only salient differences are emphasized in Appendix E.

E. Asymptotic Decay Rates

We next study the case where the false reject probability decays exponentially fast as well and thus characterize the tradeoff between the false reject exponent and the homogeneous error exponent of the misclassification error and false alarm probabilities. Recall the definition of $LD_B(\lambda, P_N, P_B)$ in (47).

Theorem 6. For any nominal distribution $P_N$ and anomalous distributions $P_T = (P_{A,1}, \ldots, P_{A,T})$, given any positive real number $\lambda \in \mathbb{R}_+$, the universal test in (51) satisfies that for each $B \in \mathcal{S}$,

$$\liminf_{n \to \infty} -\frac{1}{n} \log \beta_B(\Psi_n|P_N, P_T) \geq \lambda,$$

(63)

$$\liminf_{n \to \infty} -\frac{1}{n} \log \beta_B(\Psi_n|P_N, P_T) \geq \lambda,$$

(64)

$$\liminf_{n \to \infty} -\frac{1}{n} \log \zeta_B(\Psi_n|P_N, P_T) \geq LD_B(\lambda, P_N, P_B).$$

(65)

Conversely, for any test that ensures the homogeneous exponential decay rate of the misclassification error and false alarm is no less than $\lambda$ for all tuples of nominal and anomalous distributions, under any nominal distribution $P_N$ and anomalous distributions $P_T = (P_{A,1}, \ldots, P_{A,T})$, the false reject exponent is also upper bounded by $LD_B(\lambda, P_N, P_B)$ under each hypothesis $H_B$.

The proof of Theorem 6 requires the same modification of the proof of Theorem 4 in the manner in the proof of Theorem 3 and is thus omitted. The result in Theorem 3 follows by specializing Theorem 6 to the case of $T = 1$. Similar remarks as those for Theorem 3 apply here.

For example, the threshold $\lambda$ governs the tradeoff between the false reject exponent and the homogeneous error exponent under each hypothesis. From the definition of $LD_B(\lambda, P_N, P_B)$ in (47), it follows that the false reject exponent $LD_B(\lambda, P_N, P_B)$ in (47) decreases in $\lambda$. Similarly to the proof in Appendix H, one can show that $LD_B(\lambda, P_N, P_B) > 0$ if and only if $\lambda < GD(B, P_N, P_B)$ and the maximal false reject exponent satisfies

$$\max_{\lambda \in (0, GD(B, P_N, P_B))} LD_B(\lambda, P_N, P_B) \leq \min_{Q \in P(\mathcal{X})} \left(\sum_{i \in B} D(Q\|P_{A,i}(\cdot)) + (M - |B|)D(Q\|P_N)\right).$$

(66)

Therefore, if the threshold $\lambda < \min_{B \in \mathcal{S}} GD(B, P_N, P_B)$, then regardless of the number of outliers, the misclassification error, the false alarm and false reject probabilities decay exponentially fast for any tuple of distributions $(P_N, P_T)$ such that $\min_{B \in \mathcal{S}} GD(B, P_N, P_B)$ is strictly positive, which is almost always true for finite alphabet sequences.

IV. CONCLUSION

We revisited the universal hypothesis testing problem studied by Li et al. in [2] and derived performance guarantees for universal tests when there are potentially multiple outliers. In particular, we first study the case with at most one outlier and then generalize our results to the case where there are multiple outliers, the number of outliers is unknown and each outlier can be generated from a unique anomalous distributions. For both cases, we proposed a universal test, analyzed its performance in terms of the tradeoff among the misclassification error, false alarm and false reject probabilities. Furthermore, we demonstrate the optimality of the test in the generalized Neyman-Pearson sense proposed by Gutman [3]. Our results have brought new insights beyond [2] in several aspects, including the design of a second-order asymptotic optimal test, the dominant factors affecting performance of a test and a second-order approximation to the finite sample size approximation using recently proposed finite blocklength information theoretical tools [13], [16].

There are several avenues for future research. Firstly, it might be interesting to study tests that can ensure exponential decay misclassification error probabilities for any pair of nominal and anomalous distributions and simultaneously ensure that the false alarm and false reject probabilities are upper bounded by a constant for all pairs of nominal and anomalous distributions. Secondly, it would be interesting to study the optimality of tests in regimes other than the generalized Neyman-Pearson sense as demonstrated in this paper. For example, whether the universal tests in this paper are optimal in the finite sample regime for a set of nominal and anomalous distributions, which would be stronger that the asymptotic guarantees provided by Theorems 3 and 6. Thirdly, it would be valuable to extend our theory to the scenario where each nominal sample is generated from a
potentially different distribution in a neighborhood of a fixed distribution and then derive the performance of the optimal test, similarly to [17]. Fourthly, one might generalize our results to the case of continuous alphabet where each observed sequence is generated i.i.d. from a probability distribution function, e.g., a Gaussian distribution. Finally, it would be worthwhile to consider a sequential setting by incorporating ideas from [8] to derive second-order asymptotic limits of an optimal sequential test.

**APPENDIX**

A. Proof of Theorem 1

Recall the definitions of information densities in (2) and (3). Given any pair of distributions \((P_N, P_A)\), define the following linear combination of the third absolute moment of information densities

\[
T(P_N, P_A) := E_{P_N} \left[|_1(X|P_N, P_A) - E_{P_N}[|_1(X|P_N, P_A)]|^3\right] \\
+ (M - 2)E_{P_N} \left[|_2(X|P_N, P_A) - E_{P_N}[|_2(X|P_N, P_A)]\right].
\]

(67)

Note that \(T(P_N, P_A)\) is finite since we consider distributions \((P_N, P_A)\) with the same support on the finite alphabet \(\mathcal{X}\). Recall the definition of the scoring function \(S_i(x^n) = G_i(T_{x_1^M}, \ldots, T_{x_M^n})\) (cf. (1)) for each \(i \in [M]\). Furthermore, for any given set of sequences \(x^n = (x^n_1, \ldots, x^n_M)\), define the following two quantities

\[
i^*(x^n) := \arg\min_{i \in [M]} S_i(x^n), \quad h(x^n) := \min_{i \in [M]: i \neq i^*(x^n)} S_i(x^n).
\]

(68) (69)

Note that \(i^*(x^n)\) denotes the index of the minimal scoring function (unique with high probability as we shall show) and \(h(x^n)\) denotes the value of the second minimal value of the scoring functions. Using these two definitions, our proposed test in (13) is equivalently expressed as follows:

\[
\psi_n(x^n) = \begin{cases} 
H_i & \text{if } i^*(x^n) = i, \text{ and } h(x^n) > \lambda, \\
H_r & \text{if } h(x^n) \leq \lambda
\end{cases}
\]

(70)

We first analyze the misclassification error probabilities of our test \(\psi_n(\cdot)\) under each hypothesis. Recall that we use \(Q\) to denote a collection of \(M\) distributions \((Q_1, \ldots, Q_M)\) defined on the alphabet \(\mathcal{X}\). For any pair of nominal and anomalous distributions \((P_N, P_A)\) and for each \(i \in [M]\), we can upper bound the type-i misclassification error probability as follows:

\[
\beta_1(\psi_n|P_N, P_A) \\
= P_x(i^*(X^n) \neq i, \ h(X^n) > \lambda) \\
\leq P_x(S_i(X^n) > \lambda) \\
= \sum_{x^n \in \mathcal{X}^n: S_i(x^n) > \lambda} P^n_A(x^n) \times \left( \prod_{j \in M_i} P^n_N(x^n_j) \right) \\
= \sum_{Q \in P_n(x^n): G_i(Q) > \lambda} \sum_{x^n: \forall \ j \in [M], x^n_j \in T^n_{Q_j}} \exp \left( -n \left( D(Q_i||P_A) + H(Q_i) \right) \right) \\
\times \exp \left( -n \left( \sum_{j \in M_i} (D(Q_j||P_N) + H(Q_j)) \right) \right) \\
\leq \exp(-n\lambda) \sum_{Q \in P_n(x^n)} \sum_{x^n_j \in T^n_{Q_j}} \exp \left( -n \left( D(Q_i||P_A) \right) \right) \\
\times \exp \left( -n \left( (M - 1)D \left( \frac{\sum_{k \in M_i} Q_k}{M - 1} \left\| P_N \right\| \right) \right) \right) \leq \exp(-n\lambda) \sum_{Q \in P_n(x^n)} \sum_{x^n_j \in T^n_{Q_j}} \exp \left( -n \left( D(Q_i||P_A) \right) \right) \\
\times \exp \left( -n \left( (M - 1)D \left( \frac{\sum_{k \in M_i} Q_k}{M - 1} \left\| P_N \right\| \right) \right) \right)
\]

(71) (72) (73) (74) (75) (76) (77)
\[ \leq \exp(-n\lambda) \sum_{Q_j \in P_n(X), j \in M_i} \exp \left( -n \left( (M-1)D \left( \frac{\sum_{k \in M_i} Q_k}{M-1} \mid P_N \right) \right) \right) \] (78)

\[ \leq \exp(-n\lambda) \sum_{Q \in P^{(M-1)n}(X)} ((M-1)n + 1)^{|X|}P_N^{(M-1)n} \left( T_Q^{(M-1)n} \right) \] (79)

\[ = \exp \left( -n\lambda + |X| \log((M-1)n + 1) \right), \] (80)

where (72) follows from definitions of \( i^*(x^n) \) in (68) and \( h(x^n) \) in (69) which indicate that \( S_i(x^n) \geq h(x^n) > \lambda \) under the condition that \( i^*(x^n) \neq i \) and \( h(x^n) > \lambda \); (73) follows since under hypothesis \( H_i \), the \( i \)-th sequence \( X_i^n \) is generated i.i.d. according to the anomalous distribution \( P_A \) while all other sequences are generated i.i.d. according to the nominal distribution \( P_N \); (74) follows from the definitions of the scoring function \( S_i(\cdot) \) in (14) and \( G_i(\cdot) \) in (1) and method of types [18, Chapter 11]; (76) follows since for any sequence of distributions \( Q = (Q_1, \ldots, Q_M) \) and any distribution \( \hat{P}_N \), the following equalities hold

\[ \sum_{j \in M_i} D(Q_j||P_N) = \sum_{j \in M_i} \mathbb{E}_{Q_j} \left[ \log \frac{Q_j(X)}{P_N(X)} \right] \] (81)

\[ = \sum_{j \in M_i} \mathbb{E}_{Q_j} \left[ \log \frac{\sum_{k \in M_i} \frac{Q_k(X)}{P_N(X)}}{\sum_{k \in M_i} Q_k(X)} \right] + \log \frac{Q_j(X)}{\sum_{k \in M_i} Q_k(X)} \] (82)

\[ = \sum_{j \in M_i} \mathbb{E}_{Q_j} \left[ \log \frac{\sum_{k \in M_i} \frac{Q_k(X)}{P_N(X)}}{\sum_{k \in M_i} Q_k(X)} \right] + G_i(Q) \] (83)

\[ = (M-1)D \left( \frac{\sum_{k \in M_i} Q_k}{M-1} \mid P_N \right) + G_i(Q); \] (84)

(77) follows since the size of the type class \( |T^{(M-1)n}_{Q_j}| \leq \exp(nH(Q_j)); \) (78) follows since

\[ \sum_{Q_i \in P_n(X)} \sum_{x^n_i \in T^{(M-1)n}_{Q_j}} \exp \left( -n \left( D(Q_i||P_A) + H(Q_i) \right) \right) = \sum_{x^n_i \in X^n} P_A^n(x^n_i) = 1, \] (85)

and (79) follows from the lower bound on the probability of the type class \( T^{(M-1)n}_Q \) and the fact that summing over \( (M-1) \) concatenated types of length \( n \) is equivalent to summing over a type of length \( (M-1)n \).

Given any pair of nominal and anomalous distributions \( (P_N, P_A) \), we can upper bound the false alarm probability as follows:

\[ P_{fa}(\psi_n \mid P_N, P_A) = \mathbb{P}_t \{ h(X^n) > \lambda \} \] (86)

\[ = \sum_{i \in [M]} \mathbb{P}_t \{ i^*(X^n) = i \text{ and } h(X^n) > \lambda \} \] (87)

\[ \leq \sum_{i \in [M]} \mathbb{P}_t \{ i^*(X^n) = i \text{ and } \exists j \in M_i: S_j(X^n) > \lambda \} \] (88)

\[ \leq \sum_{i \in [M]} \sum_{j \in M_i} \mathbb{P}_t \{ S_j(X^n) > \lambda \} \] (89)

\[ \leq \sum_{i \in [M]} \sum_{j \in M_i} \sum_{x^n_j: S_j(x^n) > \lambda} \prod_{t \in [M]} P_N(x^n) \] (90)

\[ = \sum_{i \in [M]} \sum_{j \in M_i} \sum_{Q \in P_n(X)^M: x^n_j: \forall j \in [M]} \sum_{x^n_j: S_j(x^n) > \lambda} \exp \left( -n \sum_{t \in [M]} (D(Q_t||P_N) + H(Q_t)) \right) \] (91)

\[ \leq \sum_{i \in [M]} \sum_{j \in M_i} \exp(-n\lambda + |X| \log((M-1)n + 1)) \] (92)

\[ \leq M(M-1) \exp(-n\lambda + |X| \log((M-1)n + 1)), \] (93)

where (88) follows since when \( i^*(X^n) = i, h(X^n) = \min_{j \in M_i} S_j(X^n) \) and (92) follows from the steps analogously to those leading to the result in (80).

Finally, we next analyze the false reject probabilities for any \( (P_N, P_A) \). For this purpose, we need the following definition of typical sequences for each \( i \in [M]: \)

\[ T_i(P_N, P_A) := \left\{ x^n \in X^{Mn}: \forall j \in M_i, \| \hat{T}_{x^n} - P_N \|_1 \leq \sqrt{\frac{\log n}{n}} \text{ and } \| \hat{T}_{x^n} - P_A \|_1 \leq \sqrt{\frac{\log n}{n}} \right\}. \] (94)
Using Chebyshev’s inequality (c.f. [19, Lemma 24]), we conclude that for each \(i \in [M]\),
\[
\Pr \{ X^n \notin T_i(P_N, P_A) \} \leq \frac{2M|\mathcal{X}|}{n^2} := \mu_n. \tag{95}
\]

In subsequent analysis, we need to use the following properties of \(G_i(Q) \) (cf. (1)) for each \(i \in [M]\) and any given vector of distributions \(Q = (Q_1, \ldots, Q_M) \in (\mathcal{P}(\mathcal{X}))^M\),
\[
\frac{\partial G_i(Q)}{\partial Q_j(x)} = \log \left( \frac{M - 1}{Q_j(x)} \right) + \sum_{k \in M, k \neq j} Q_k(x), \quad j \in \mathcal{M}_i, \quad x \in \text{supp}(Q_j), \tag{96}
\]
\[
\frac{\partial^2 G_i(Q)}{\partial Q_j(x) Q_l(x)} = \frac{\sum_{k \in M, k \neq j} Q_k(x) - Q_j(x)}{Q_j(x)} \left( \sum_{k \in M, k \neq j} Q_k(x) \right), \quad j \in \mathcal{M}_i, \quad x \in \text{supp}(Q_j), \tag{97}
\]
\[
\frac{\partial^2 G_i(Q)}{\partial Q_j(x) Q_l(x)} = \frac{1}{\sum_{k \in M} Q_k(x)} (j, l) \in \mathcal{M}_i \times \mathcal{M}_i, \quad x \in \text{supp}(Q_j) \cap \text{supp}(Q_l). \tag{98}
\]

For each \(i \in [M]\), define the vector of distributions \(P_i := (Q_1, \ldots, Q_M) \) with \(Q_i = P_A\) and \(Q_j = P_N\) for all \(j \in \mathcal{M}_i\). Under hypothesis \(H_i\), given any set of \(M\) sequences \(x^n \in T_i(P_N, P_A)\), since the KL divergence is continuous, one can apply a Taylor expansion of \(G_j(\hat{T}_{x_1^n}, \ldots, \hat{T}_{x_M^n}) \) (cf. (14)) around \(P_i\), for each \(j \in \mathcal{M}_i\), we have
\[
G_j(\hat{T}_{x_1^n}, \ldots, \hat{T}_{x_M^n}) = D\left( P_A \left\| \frac{(M - 2)P_N + P_A}{M - 1} \right\| \right) + \sum_{x \in \mathcal{X}} (\hat{T}_{x_1^n}(x) - P_A(x)I_1(x|P_N, P_A) \right)
\]
\[
+ \sum_{l \in \mathcal{M}_i, j \neq l} \left( D\left( P_N \left\| \frac{(M - 2)P_N + P_A}{M - 1} \right\| \right) + \sum_{x \in \mathcal{X}} (\hat{T}_{x_1^n}(x) - P_N(x)I_2(x|P_N, P_A) \right)
\]
\[
+ \sum_{l \in [M]} O(\|\hat{T}_{x^n} - P_A\|^2) \right) \tag{99}
\]
\[
= \frac{1}{n} \sum_{t \in [n]} \left( I_1(x_{i,t}|P_N, P_A) + \sum_{l \in \mathcal{M}_i, j \neq l} I_2(x_{i,t}|P_N, P_A) \right) + O\left( \frac{\log n}{n} \right), \tag{100}
\]
and for \(j = i\),
\[
G_j(\hat{T}_{x_1^n}, \ldots, \hat{T}_{x_M^n}) = O\left( \frac{\log n}{n} \right). \tag{101}
\]

With the above definitions and results, we can now upper bound the false reject probability of our test (cf. (13)) under each hypothesis \(H_i\) with \(i \in [M]\) with respect to any pair of distributions \((P_N, P_A)\) as follows:
\[
\zeta_i(\psi_t|P_N, P_A) = \Pr \{ h(X^n) \leq \lambda \}
\leq \Pr \{ \min_{j \in \mathcal{M}_i} G_j(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}) \leq \lambda \}
= 1 - \Pr \{ \forall j \in \mathcal{M}_i, \quad G_j(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}) > \lambda \}, \tag{102}
\]
where (103) follows since \(h(X^n) \geq \min_{j \in \mathcal{M}_i} G_j(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n})\), which is implied by the definition of \(h(x^n)\) in (69).

For simplicity, given random variables \(X_1, \ldots, X_M\), for each \(i \in [M]\) and \(j \in \mathcal{M}_i\), define the information density
\[
t_{i,j}(X_1, \ldots, X_M|P_N, P_A) := t_{1}(X_i|P_N, P_A) + \sum_{l \in \mathcal{M}_i, j \neq l} t_{2}(X_l|P_N, P_A), \tag{103}
\]
and for each \(t \in [n]\), we use \(X_t\) to denote the snapshot of the \(M\) sequences at time \(t\), i.e., \(X_{1,t}, \ldots, X_{M,t}\).

The second term in (104) can be lower bounded as follows:
\[
\Pr \{ \forall j \in \mathcal{M}_i, \quad G_j(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}) > \lambda \}
\geq \Pr \{ \forall j \in \mathcal{M}_i, \quad G_j(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}) > \lambda \ \text{and} \ \ X^n \in T_i(P_N, P_A) \} \tag{105}
\]
\[
\geq \Pr \{ \forall j \in \mathcal{M}_i, \quad \frac{1}{n} \sum_{t \in [n]} t_{i,j}(X_t|P_N, P_A) > \lambda + O\left( \frac{\log n}{n} \right) \ \text{and} \ \ X^n \in T_i(P_N, P_A) \} \tag{106}
\]
\[
\geq \Pr \{ \forall j \in \mathcal{M}_i, \quad \frac{1}{n} \sum_{t \in [n]} t_{i,j}(X_t|P_N, P_A) > \lambda + O\left( \frac{\log n}{n} \right) \} - \mu_n, \tag{107}
\]
where (107) follows from the result in (95) and the Taylor expansion in (100), and (108) follows from the result in (95).

Recall that under $P_i$, for each $t \in [n]$, $X_t = (X_{1,t}, \ldots, X_{M_t})$ are independent where $X_{1,t} \sim P_A$ and $X_{j,t} \sim P_N$ for $j \in M_i$. Recalling definitions of $GD_M(P_N, P_A)$ in (4), $V_M(P_N, P_A)$ in (5) and $COV_M(P_N, P_A)$ in (6), we have that for any $i \in [M]$ and $j \in M_i$

$$E_{P_i}[i_{i,j}(X_t|P_N, P_A)] = GD_M(P_N, P_A),$$

$$\text{Var}_{P_i}[i_{i,j}(X_t|P_N, P_A)] = V_M(P_N, P_A),$$

and for any $k \in M_{i,j}$, the covariance of $(i_{i,j}(X_t|P_N, P_A), i_{i,k}(X_t|P_N, P_A))$ satisfies

$$COV_{P_i}[i_{i,j}(X_t|P_N, P_A), i_{i,k}(X_t|P_N, P_A)] = COV_M(P_N, P_A),$$

where the justification of (111) is provided in Appendix 1.

Recall the definition of $V_M(P_N, P_A)$ in (7). Applying the multivariate Berry-Esseen theorem [20], the first term in (108) is bounded below as follows:

$$P_i\left\{ \forall j \in M_i, \frac{1}{n} \sum_{t \in [n]} i_{i,j}(X_t|P_N, P_A) > \lambda + O\left(\frac{\log n}{n}\right) \right\} \geq Q_{M-1}\left(\sqrt{n}\left(\lambda - GD_M(P_N, P_A) + O\left(\frac{\log n}{n}\right)\right) \times 1_{M-1}; V_M(P_N, P_A) \right) + O\left(\frac{1}{\sqrt{n}}\right),$$

where $Q_{M-1}(\cdot)$ is the multivariate generalization of the complementary Gaussian cdf defined in (8).

Using (104) and (112), we have that for any $(P_N, P_A)$, the false reject probability is upper bounded as follows:

$$\zeta_i(P_N, P_A) \leq 1 - Q_{M-1}\left(\lambda - GD_M(P_N, P_A) + O\left(\frac{\log n}{n}\right) \times 1_{M-1}; V_M(P_N, P_A) \right) + O\left(\frac{1}{\sqrt{n}}\right).$$

### B. Proof of Theorem 2

Note that the converse proof without a constraint on the false alarm probability is also a converse proof with a constraint on the false alarm probability. Therefore, in the subsequent proof, we drop the constraint on the false alarm probability and focus on the misclassification and the false reject probabilities.

We first relate the performances of any test with the type-based test (i.e., a test which uses only the types (empirical distributions) of the sequences $(\hat{T}_{X_i^T}, \ldots, \hat{T}_{X_{N_i}^T})$), as demonstrated in the following lemma.

**Lemma 1.** Given any test $\phi_n$, for any $\kappa \in [0, 1]$, we can construct a type-based test $\phi_n^T$ such that for each $i \in [M]$ and any pair of distributions $(P_N, P_A)$,

$$\beta_i(\phi_n|P_N, P_A) \geq \frac{1 - \kappa}{M-1} \beta_i(\phi_n^T|P_N, P_A),$$

$$\zeta_i(P_N, P_A) \geq \kappa \zeta_i(\phi_n^T|P_N, P_A).$$

The proof of Lemma 1 is inspired by [3, Lemma 2] and [4, Lemma 5.1] and provided in Appendix F.

We then show that for any type-based test, if we require that the misclassification error probabilities under each hypothesis decay exponentially fast for all pairs of distributions, then the false reject probability under each hypothesis for any particular pair of distributions can be lower bounded by an information spectrum bound, which is the cdf of the second minimal values of the scoring functions $\{G_i(\hat{T}_{X_i^T}, \ldots, \hat{T}_{X_{N_i}^T})\}$.

For simplicity, let

$$\eta_{n,M} := \frac{M|X| \log(n+1)}{n}. \quad (116)$$

Furthermore, given any tuple of types $Q = (Q_1, \ldots, Q_M) \in (P_n(X))^M$ and any $\lambda \in \mathbb{R}_+$, let

$$g^*(Q) := \min_{i \in [M]} G_i(Q),$$

$$g(Q) := \min_{i \in [M]: G_i(Q) > g^*(Q)} G_i(Q) \quad (118)$$

denote the minimal and second minimal values of $\{G_i(Q)\}_{i \in [M]}$.

**Lemma 2.** Given any $\lambda \in \mathbb{R}_+$, for any type-based test $\phi_n^T$ such that for all pair of distributions $(\hat{P}_N, \hat{P}_A)$,

$$\max_{i \in [M]} \beta_i(\phi_n^T|\hat{P}_N, \hat{P}_A) \leq \exp(-n\lambda),$$

\(119\)
then for any pair of distributions \((P_N, P_A)\) and for each \(i \in [M]\), we have
\[
\zeta_i(\phi_n^T|P_N, P_A) \geq \mathbb{P}_i\left\{g(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}) + \eta_{n,M} \leq \lambda\right\}. \tag{120}
\]

The proof of Lemma 2 is provided in Appendix G.

Combining Lemmas 1 and 2 with \(\kappa = 1 - \frac{1}{n}\) and noting that \(g(\hat{T}_{x_{i1}}, \ldots, \hat{T}_{x_{iM}}) = h(x^n)\) (cf. (69)) for any \(x^n = (x_1^n, \ldots, x_M^n)\), we obtain the following corollary.

**Corollary 1.** Given any \(\lambda \in \mathbb{R}_+\), for any test \(\phi_n\) satisfying that for all pairs of distributions \((\hat{P}_N, \hat{P}_A)\)
\[
\max_{i \in [M]} \beta_i(\phi_n|\hat{P}_N, \hat{P}_A) \leq \exp(-n\lambda), \tag{121}
\]
we have that for any pair of distributions \((P_N, P_A)\) and for each \(i \in [M]\)
\[
\zeta_i(\phi_n|P_N, P_A) \geq \left(1 - \frac{1}{n}\right) \mathbb{P}_i\left\{h(X^n) + \eta_{n,M} + \frac{\log n + \log(M-1)}{n} \leq \lambda\right\}. \tag{122}
\]

Using Corollary 1, with any test \(\phi_n\) satisfying (121), for any pair of distributions \((P_N, P_A)\), we have that for each \(i \in [M]\) and any \(j \in M_i\),
\[
\zeta_i(\phi_n|P_N, P_A) \geq \left(1 - \frac{1}{n}\right) \mathbb{P}_i\left\{h(X^n) + \eta_{n,M} + \frac{\log n + \log(M-1)}{n} \leq \lambda, \right\} \tag{123}
\]
\[
\zeta_i(\phi_n|P_N, P_A) \geq \left(1 - \frac{1}{n}\right) \mathbb{P}_i\left\{\min_{j \in M_i} G_j(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}) + \eta_{n,M} + \frac{\log n + \log(M-1)}{n} \leq \lambda, \right\} \tag{124}
\]

We first focus on the second term in the bracket of (124). Given any \(i \in [M]\), we have that for each \(j \in M_i:\)
\[
\mathbb{P}_i\{G_j(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}) \leq G_i(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}})\}
\]
\[
\leq \mathbb{P}_i\{G_j(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}) \leq G_i(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}), X^n \in T_i(P_N, P_A)\} + \mathbb{P}_i\{X^n \notin T_i(P_N, P_A)\} \tag{125}
\]
\[
\leq \mathbb{P}_i\left\{\frac{1}{n} \sum_{i \in [n]} \left(\eta_1(X_{i,t}|P_N, P_A) + \sum_{l \in M_{i,j}} \eta_2(X_{i,t}|P_N, P_A)\right) \leq O\left(\frac{\log n}{n}\right)\right\} + \mu_n \tag{126}
\]
\[
\leq Q\left(\frac{\sqrt{n}(G_M(P_N, P_A) + O(\frac{\log n}{n}))}{\sqrt{V_M(P_N, P_A)}}\right) + \frac{6T(P_N, P_A)}{\sqrt{n(V_M(P_N, P_A))^3}} + \mu_n \tag{127}
\]
\[
\leq \exp\left(-\frac{n(G_M(P_N, P_A) + O(\frac{\log n}{n}))^2}{2V_M(P_N, P_A)}\right) + \frac{6T(P_N, P_A)}{\sqrt{n(V_M(P_N, P_A))^3}} + \mu_n \tag{128}
\]
\[
=: \kappa_n = O\left(\frac{1}{\sqrt{n}}\right), \tag{129}
\]
where (126) follows from Taylor expansions in (100) and (101) and the upper bound on the atypical set in (95), (127) follows from the Berry-Esseen theorem [21], [22] and (128) follows since \(Q(x) \leq \exp\left(-\frac{x^2}{2}\right)\) for any \(x > 0\). Therefore, we conclude that for each \(i \in [M]\),
\[
\mathbb{P}_i\{h(X^n) \neq \min_{j \in M_i} G_j(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}})\}
\]
\[
= \mathbb{P}_i\{\exists j \in M_i \text{ s.t. } G_j(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}) < G_i(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}})\} \tag{130}
\]
\[
\leq \sum_{j \in M_i} \mathbb{P}_i\{G_j(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}}) < G_i(\hat{T}_{X_{i1}}, \ldots, \hat{T}_{X_{iM}})\} \tag{131}
\]
\[
\leq 1 - (M-1)\kappa_n. \tag{132}
\]
Finally, analogously to the achievability proof, we analyze the first term in the bracket of (124):
\[
\begin{align*}
\mathbb{P}_i \left\{ \min_{j \in M_i} G_j(T_{X_n}, \ldots, T_{X_{n,M}}) + \eta_{n,M} + \frac{\log n + \log (M-1)}{n} \leq \lambda \right\} \\
= 1 - \mathbb{P}_i \left\{ \forall j \in M_i, \ G_j(T_{X_n}, \ldots, T_{X_{n,M}}) + \eta_{n,M} + \frac{\log n + \log (M-1)}{n} > \lambda \right\} \\
\geq 1 - \mathbb{P}_i \left\{ X^n \notin T_i(P_N, P_A) \right\} \\
- \mathbb{P}_i \left\{ \forall j \in M_i, \ G_j(T_{X_n}, \ldots, T_{X_{n,M}}) + \eta_{n,M} + \frac{\log n + \log (M-1)}{n} > \lambda \right\} \quad \text{(133)}
\end{align*}
\]

where in (134), the definition of the typical set \(T_i(P_N, P_A)\) was in (94), (135) follows from the result in (95) that upper bounds the probability of \(\mathbb{P}_i \left\{ X^n \notin T_i(P_N, P_A) \right\}\), the Taylor expansion of \(G_j(T_{X_n}, \ldots, T_{X_{n,M}})\) exactly the same as in (108) and the fact that \(\eta_{n,M} = O(\log n/n)\), and (136) follows from the multivariate Berry-Esséen theorem similarly to (112). Combining (124) and (136), we conclude that
\[
\min_{i \in [M]} \zeta_i(\phi_n|P_N, P_A) \geq 1 - \sum_{M-1} \left( \sqrt{n} \left( \lambda - GD_M(P_N, P_A) + O\left(\frac{\log n}{n}\right) \right) \times 1_{M-1} + 0_{M-1}; V_M(P_N, P_A) \right) + O\left(\frac{1}{\sqrt{n}}\right). 
\]

The proof of Theorem 2 is now completed.

C. Proof of Theorem 3

1) Achievability: We make use the same test \(\phi_n(\cdot)\) (cf. (13) and (70)) as in the achievability proof of Theorem 1.

The analyses of the misclassification error probabilities \(\beta_i(\psi_n|P_N, P_A)\) and the false alarm probability \(P_{\text{fa}}(\psi_n|P_N, P_A)\) are exactly the same as in Appendix A. It suffices to bound the false reject probability of our test for a particular pair of distributions \((P_N, P_A)\). For each \(i \in [M]\), we have that
\[
\zeta_i(\psi_n|P_N, P_A) = \mathbb{P}_i \{ \phi_n(X^n) = H_r \} = \mathbb{P}_i \{ h(X^n) \leq \lambda \} = \mathbb{P}_i \{ 3 (j, k) \in [M]^2 \text{ s.t. } j \neq k, \ S_j(X^n) \leq \lambda \text{ and } S_k(X^n) \leq \lambda \} \leq \sum_{(j, k) \in [M]^2 \text{ s.t. } j \neq k} \mathbb{P}_i \{ S_j(X^n) \leq \lambda \text{ and } S_k(X^n) \leq \lambda \} \leq \frac{M(M-1)}{2} \max_{(j, k) \in [M]^2 \text{ s.t. } j \neq k} \mathbb{P}_i \{ S_j(X^n) \leq \lambda \text{ and } S_k(X^n) \leq \lambda \}. 
\]

We now focus on upper bound the probability term in (142). For any \((j, k) \in [M]^2\), given any \(i \in [M]\), we have
\[
\begin{align*}
\mathbb{P}_i \{ S_j(X^n) \leq \lambda \text{ and } S_k(X^n) \leq \lambda \} \leq \sum_{X^n \in X^{Mn}: S_j(X^n) \leq \lambda \text{ and } S_k(X^n) \leq \lambda} \mathbb{P}_i(X^n) \leq \sum_{Q \in (P_n(X))^M, G_j(Q) \leq \lambda, G_k(Q) \leq \lambda} \exp \left\{ -n \left( D(Q_i\|P_A) + \sum_{l \in M_i} D(Q_j\|P_N) \right) \right\} \leq \sum_{Q \in (P_n(X))^M, G_j(Q) \leq \lambda, G_k(Q) \leq \lambda} \exp \left\{ -n \min_{Q \in (P_n(X))^M, G_j(Q) \leq \lambda, G_k(Q) \leq \lambda} \left( D(Q_i\|P_A) + \sum_{l \in M_i} D(Q_j\|P_N) \right) \right\} \leq (n+1)^M |X|^n \exp \left\{ -n \min_{Q \in (P_n(X))^M, G_j(Q) \leq \lambda, G_k(Q) \leq \lambda} \left( D(Q_i\|P_A) + \sum_{l \in M_i} D(Q_j\|P_N) \right) \right\}. 
\end{align*}
\]

Combining (142), (146) and using the definitions of \(LD_i(\cdot)\) in (9), we have that for each \(i \in [M]\) and any pair of distributions \((P_N, P_A)\), the \(i\)-th false reject probability satisfies for any \(\lambda \in \mathbb{R}_+\)
\[
\liminf_{n \to \infty} \frac{-1}{n} \log \zeta_i(\psi_n|P_N, P_A) \geq LD_i(\lambda|P_N, P_A). 
\]
2) **Converse:** For simplicity, let

\[ \kappa_{n,M} := \eta_{n,M} + \frac{\log n + \log(M - 1)}{n}. \]  

(148)

Using Corollary 1, we have that for any test \( \phi_n \) such that the misclassification error probabilities decay exponentially fast with speed at least \( \lambda \) for all pairs of distributions, given any \((P_N, P_A)\), for each \( i \in [M] \), the \( i \)-th false reject probability \( \zeta_i(\phi_n|P_N, P_A) \) satisfies

\[ \left( 1 - \frac{1}{n} \right) \times \zeta_i(\phi_n|P_N, P_A) \]

\[ \geq P_i\{h(X^n) + \kappa_{n,M} \leq \lambda \} \]

\[ = P_i\{\exists (j, k) \in [M]^2: j \neq k, S_j(X^n) + \kappa_{n,M} \leq \lambda \text{ and } S_k(X^n) + \kappa_{n,M} \leq \lambda \} \]

\[ \geq \max_{(j, k) \in [M]^2: j \neq k} P_i\{S_j(X^n) + \kappa_{n,M} \leq \lambda \text{ and } S_k(X^n) + \kappa_{n,M} \leq \lambda \} \]

\[ \geq (n + 1)^\{-M|\mathcal{X}|\} \max_{(j, k) \in [M]^2: j \neq k} \sum_{Q \in (P_{n}(X))^M} \exp\left( -n \min_{Q \in (P_{n}(X))^M, G_i(Q) \leq \lambda - \kappa_{n,M}} \left( D(Q_i\|P_A) + \sum_{i \in M_i} D(Q_j\|P_N) \right) \right), \]

(152)

where (150) follows from the definition of \( h(x^n) \) in (69) and (152) follows similarly to (145).

Using the continuity of \((P_N, P_A)\) to \( LD_i(\lambda|P_N, P_A) \) (cf. (9)) for any \( \lambda \in \mathbb{R}_+ \), the definition of \( \kappa_{n,M} \) in (148) and the results in (152), we have that for each \( i \in [M] \),

\[ \limsup_{n \to \infty} -\frac{1}{n} \log \zeta_i(\phi_n|P_N, P_A) \leq LD_i(\lambda|P_N, P_A) \]

(153)

for any test \( \phi_n \) satisfying (33).

**D. Proof of Theorem 4**

The proof of Theorem 4 is a generalization of Theorem 1 and thus we only emphasize the differences.

For subsequent analyses, define the following linear combination of third absolute moments

\[ T(B, C, P_N, P_B) := \sum_{i \in (B \cap \mathcal{M}_C)} \mathbb{E}_{P_{\lambda, \beta(i)}} \left[ |t_{1, \beta(i)}(X|B, C, P_N, P_B)|^3 \right] \]

\[ + \sum_{i \in (M \setminus B \cap \mathcal{M}_C)} \mathbb{E}_{P_N} \left[ |t_2(X|B, C, P_N, P_B)|^3 \right]. \]

(154)

Note that \( T(B, C, P_N, P_B) \) is finite since we consider distributions \((P_N, P_B)\) with the same support on the finite alphabet \( \mathcal{X} \).

1) **Achievability:** Recall that \( \mathcal{S} = \bigcup_{i \in [M]} \mathcal{S} \) denotes all possible subsets of \([M]\) with at most \( \lceil \frac{M}{2} \rceil - 1 \) elements. For any \( C \in \mathcal{S} \), recall the definition of the scoring function \( S_C(x^n) \) in (52). Recall the definition of the scoring function \( S_B(x^n) = G_B(\hat{T}_x^n, \ldots, \hat{T}_{x_M^n}) \) for any \( B \in \mathcal{S} \). Given any \( x^n \), parallel to (68) and (69), define two quantities

\[ \mathcal{I}^*(x^n) := \arg\min_{B \in \mathcal{S}} S_B(x^n), \]

\[ h_\mathcal{S}(x^n) := \min_{B \in \mathcal{S}, B \neq \mathcal{I}^*(x^n)} S_B(x^n). \]

(155)

(156)

Note that \( \mathcal{I}^*(x^n) \) denotes the set \( B \) that minimizes the scoring function and \( h_\mathcal{S}(x^n) \) denotes the second minimal value of the scoring function.

The universal rest in (51) is equivalently expressed as follows:

\[ \Psi_n(x^n) = \begin{cases} H_B & \text{if } \mathcal{I}^*(x^n) = B, \text{ and } h_\mathcal{S}(x^n) > \lambda \vspace{1em} \\ H_r & \text{if } h_\mathcal{S}(x^n) \leq \lambda. \end{cases} \]

(157)
We then analyze the performance of the test in (157). We first analyze the misclassification error probability. Given any $B \in S$, under any tuple of distributions $P_B = (P_N, P_{A,1}, \ldots, P_{A,|B|})$, similarly to the case with at most one outlier, the misclassification error is upper bounded as follows:

\begin{align*}
\beta_B(\Psi_n|P_N, P_T) &= \mathbb{P}_B\{T^*(X^n) \neq B \text{ and } h_S(X^n) > \lambda\} \\
&\leq \mathbb{P}_B\{S_B(X^n) > \lambda\} \\
&= \sum_{B \in S} \sum_{x^n: x^n \in T^*_B} \left( \prod_{i \in B} P_{A_{B(i)}(x^n_i)} \right) \times \left( \prod_{j \in M_B} P_N(x^n_j) \right) \\
&= \sum_{B \in S} \sum_{x^n: x^n \in T^*_B} \exp \left( -n \left( \sum_{i \in B} D(Q_i||P_{A_{B(i)}}) + \sum_{j \in M_B} D(Q_i||P_N) + \sum_{i \in [M]} H(Q_i) \right) \right) \\
&\leq \exp(-n\lambda) \sum_{Q_i \in \mathcal{P}_n(\lambda), j \in M_B} \exp \left( -n(M - |B|)D \left( \frac{\sum_{k \in M_B} Q_k}{M-T} \bigg| P_N \right) \right) \\
&\leq \exp \left( -n\lambda + |\mathcal{X}| \log((M - |B|)n + 1) \right) \\
&\leq \exp \left( -n\lambda + |\mathcal{X}| \log((M - 1)n + 1) \right). 
\end{align*}

We then analyze the false alarm probability. Given any nominal distribution $P_N$, the false alarm probability is upper bounded as follows:

\begin{align*}
P_{fa}(\Psi_n|P_N, P_T) &:= \mathbb{P}_T\{h_S(X^n) > \lambda\} \\
&= \sum_{B \in S} \mathbb{P}_T\{T^*(X^n) = B \text{ and } h_S(X^n) > \lambda\} \\
&\leq \sum_{B \in S} \mathbb{P}_T\{T^*(X^n) = B \text{ and } \exists C \in S : C \neq B, S_C(X^n) > \lambda\} \\
&\leq \sum_{B \in S} \sum_{C \in S : C \neq B} \mathbb{P}_T\{S_C(X^n) > \lambda\} \\
&\leq \sum_{B \in S} \sum_{C \in S : C \neq B} \exp \left( -n\lambda + |\mathcal{X}| \log((M - 1)n + 1) \right) \\
&\leq |S|^2 \exp \left( -n\lambda + |\mathcal{X}| \log((M - 1)n + 1) \right). 
\end{align*}

where (170) follows from similar steps leading to (164).

Finally, we analyze the false reject probability of the tests. For this purpose, we need a generalized version of the typical set in (94). For each $B \in S$ and any $P_B$, define

\[ T_B(P_B) := \left\{ x^n \in \mathcal{X}^M : \forall j \in B, \| \hat{T}_j - P_{A_{B(j)}} \|_{\infty} \leq \sqrt{\log n \over n}, \right. \]

\[ \left. \quad \text{and } \forall j \in M_B, \| \hat{T}_j - P_N \|_{\infty} \leq \sqrt{\log n \over n} \right\}. \]

Similarly to (95), for each $B \in S$, we have

\[ \mathbb{P}_B\{X^n \notin T_B(P_B)\} \leq \frac{2M|\mathcal{X}|}{n^2}. \]

Recall the definitions of the mixture distribution in (39) and the information densities in (40) and (41). Under each hypothesis $H_B$, given any observed sequences $x^n \in T_B(P_B)$, applying Taylor expansions of $G_C(T_{x^n_1}, \ldots, T_{x^n_M})$ for $C \in S$ around $P_B$ yields...
• if $C \neq B$, then

$$G_C(\hat{T}_{x^*_1}, \ldots, \hat{T}_{x^*_n}) = \sum_{j \in (B^c \cap M_c)} \left( D(P_{A,B(j)} \| P_{\text{Mix}}^{B,C,P_N,P_B}) + \sum_x (\hat{T}_{x^*_j}(x) - P_{A,B(j)}(x)) t_{1,B,j}(x|B,C,P_N,P_B) + O\left( \|\hat{T}_{x^*_j} - P_{A,B(j)}\|^2 \right) \right) + \sum_{j \in (B \cap M_B \cap M_c)} \left( D(P_N \| P_{\text{Mix}}^{B,C,P_N,P_B}) + \sum_x (\hat{T}_{x^*_j}(x) - P_N(x)) t_{2}(x|B,B,P_B) + O\left( \|\hat{T}_{x^*_j} - P_N\|^2 \right) \right) = \frac{1}{n} \sum_{t \in [n]} \left( \sum_{j \in (B^c \cap M_c)} t_{1,B,j}(x_{j,t}|B,C,P_N,P_B) + \sum_{j \in (B \cap M_B \cap M_c)} t_{2}(x_{j,t}) \right) + O\left( \frac{\log n}{n} \right); \quad (175)$$

• if $C = B$, then

$$G_C(\hat{T}_{x^*_1}, \ldots, \hat{T}_{x^*_n}) = O\left( \frac{\log n}{n} \right). \quad (176)$$

The false reject probability is then upper bounded as follows:

$$\zeta_B(\Psi_n|P_N,P_T) = P_B\{h_B(X^n) \leq \lambda\} \quad (177)$$

$$\leq P_B\left\{ \min_{C \in S_B} G_C(\hat{T}_{X^*_1}, \ldots, \hat{T}_{X^*_n}) \leq \lambda \right\} \quad (178)$$

$$= 1 - P_B\left\{ \forall C \in S_B : G_C(\hat{T}_{X^*_1}, \ldots, \hat{T}_{X^*_n}) > \lambda \right\}, \quad (179)$$

where $S_B = \{C \in S_B\}$ denotes the set of sets in $S$ that are not equal to $B$. We now analyze the probability term in (179). Recall that given any $(B,C) \in S^2$ and any variables $(X_1, \ldots, X_M)$,

$$t_{B,C}(X_1, \ldots, X_M|P_N,P_B) = \sum_{j \in (B \cap M_B \cap M_c)} t_{1,B,j}(X_j|B,C,P_N,P_B) + \sum_{j \in (M_B \cap M_c)} t_{2}(X_j|B,C,P_N,P_B). \quad (180)$$

For each $t \in [n]$, we use $X_t$ to denote $X_{1,t}, \ldots, X_{M,t}$. Using Taylor expansions in (175), (176), for any $B \in S$,

$$P_B\{ \forall C \in S_B : G_C(\hat{T}_{X^*_1}, \ldots, \hat{T}_{X^*_n}) > \lambda \} \quad (181)$$

$$\geq P_B\{ \forall C \in S_B : G_C(\hat{T}_{X^*_1}, \ldots, \hat{T}_{X^*_n}) > \lambda, X^n \in T_B(P_B) \}$$

$$\geq P_B\left\{ \forall C \in S_B : \frac{1}{n} \sum_{t \in [n]} t_{B,C}(X_t|P_N,P_B) > \lambda + O\left( \frac{\log n}{n} \right) \right\} - P_B\{X^n \notin T_B(P_B)\} \quad (182)$$

$$= P_B\left\{ \forall C \in S_B : \frac{1}{n} \sum_{t \in [n]} t_{B,C}(X_t|P_N,P_B) > \lambda + O\left( \frac{\log n}{n} \right) \right\} - \frac{2M|X|}{n^2}. \quad (183)$$

Note that $S_B$ denotes all subsets of $[M]$ with size no greater than $T$ excluding the set $B$ thus each element in $S_B$ is a subset of $[M]$. There are in total $|S| - 1$ elements in the set $S_B$ and there are $|S| - 1$ inequalities (183) that need to be satisfied simultaneously. Recall that the elements in $S_B$ are ordered as $\{C_1, \ldots, C_{|S| - 1}\}$. This way, the probability term in (183) is equivalent to

$$P_B\left\{ \forall i \in [|S| - 1] : \frac{1}{n} \sum_{t \in [n]} t_{B,C}(X_t|P_N,P_B) > \lambda + O\left( \frac{\log n}{n} \right) \right\}. \quad (184)$$

Recall the definitions of $GD(B,C,P_N,P_B)$ in (43), $V(B,C,P_N,P_B)$ in (44). Note that for each $i \in [|S| - 1]$ and each $t \in [n]$,

$$E_{P_B}[t_{B,C}(X_t|P_N,P_B)] = GD(B,C_i,P_N,P_B). \quad (185)$$

$$\text{Var}_{P_B}[t_{B,C}(X_t|P_N,P_B)] = V(B,C_i,P_N,P_B). \quad (186)$$

Furthermore, for any $(i,k) \in [|S| - 1]^2$ such that $i \neq k$, we have

$$\text{Cov}_{P_B}(t_{B,C}(X_i|P_B), t_{B,D}(X_i|P_B)) = \text{Cov}(C_i, C_j, P_B). \quad (187)$$

Using (183) to (187), and applying the multivariate Berry-Esseen theorem similarly to (112), we have

$$P_B\{ \forall C \in S_B : G_C(\hat{T}_{X^*_1}, \ldots, \hat{T}_{X^*_n}) > \lambda \} \quad (188)$$

$$\geq Q_{|S| - 1}\left( \sqrt{n\tilde{\mu}}(\lambda, P_N,P_B); 0_{|S| - 1}; V(B,P_N,P_B) \right) + O\left( \frac{1}{\sqrt{n}} \right), \quad (188)$$
where $\mu(\lambda, P_N, P_B)$ denotes the vector $(\lambda - \text{GD}(B, C_1, P_N, P_B) + O(\log n/n), \ldots, \lambda - \text{GD}(B, C_{|S|-1}, P_N, P_B) + O(\log n/n))$. Combining (179) and (188), we conclude that for any $B \in S$, the false reject probability under hypothesis $H_B$ satisfies

$$\zeta_B(\Psi_n|P_N, P_T) \leq 1 - Q_{|S|-1}(\sqrt{n}\mu(\lambda, P_N, P_B); 0_{|S|-1}; 0_{|S|-1}; V(B, P_N, P_B)) + O\left(\frac{1}{\sqrt{n}}\right).$$  

(189)

The proof of Theorem 4 is completed.

E. Proof of Theorem 5

Recall the definition of $h_T(x^n)$ in (156) and $\eta_{n,M}$ in (116). For ease of notation, let

$$\eta_{n,M,T} := \eta_{n,M} + \frac{\log n + \log(|S|)}{n}.$$  

(190)

The following corollary is key to the converse proof of Theorem 4.

**Corollary 2.** Given any $\lambda \in \mathbb{R}_+$, for any test $\phi_n$ such that for all tuples of nominal and anomalous distributions $(\bar{P}_N, \bar{P}_T)$,

$$\beta_B(\phi_n|\bar{P}_N, \bar{P}_T) \leq \exp(-n\lambda),$$  

(191)

then for any tuple of nominal and anomalous distributions $(P_N, P_T)$, for each $B \in S$,

$$\zeta_B(\phi_n|P_N, P_T) \geq \left(1 - \frac{1}{n}\right) P_B\{h_S(x^n) + \eta_{n,M,T} \leq \lambda\}.$$  

(192)

The proof of Corollary 2 is similar to that of Corollary 1 and is thus omitted.

Using Corollary 2, for any test $\phi_n$ satisfying (191), given any tuple of distributions $(P_N, P_T)$, for each $B \in S$, the false reject probability is lower bounded by

$$\zeta_B(\phi_n|P_N, P_T) \geq \left(1 - \frac{1}{n}\right) P_B\{h_S(x^n) + \eta_{n,M,T} \leq \lambda\} \geq \left(1 - \frac{1}{n}\right) \min_{\psi \in S_B} G_C(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}, \hat{T}_{X_{|S|-1}^n}, \eta_{n,M,T} \leq \lambda) - P_B\{h_S(x^n) \neq \min_{\psi \in S_B} G_C(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}, \hat{T}_{X_{|S|-1}^n})\}.$$  

(193)

$$\geq \left(1 - \frac{1}{n}\right) P_B\{\min_{\psi \in S_B} G_C(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}, \hat{T}_{X_{|S|-1}^n}) \geq \eta_{n,M,T} \leq \lambda\} + O\left(\frac{1}{\sqrt{n}}\right),$$  

(194)

where (195) is justified in Appendix J.

Analogous to (188), using the multivariate Berry-Esseen theorem, we conclude that

$$P_B\{\min_{\psi \in S_B} G_C(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}) \geq \eta_{n,M,T} \leq \lambda\} = 1 - P_B\{\forall \psi \in S_B, G_C(\hat{T}_{X_1^n}, \ldots, \hat{T}_{X_M^n}) > \eta_{n,M,T} \geq \lambda\} \geq 1 - Q_{|S|-1}(\sqrt{n}\mu(\lambda, P_N, P_B); 0_{|S|-1}; 0_{|S|-1}; \eta_{n,M,T} \leq \lambda) + O\left(\frac{1}{\sqrt{n}}\right).$$  

(196)

The proof of Theorem 5 is thus completed by combining (195) and (197).

F. Proof of Lemma 1

For simplicity, let $Q := (Q_1, \ldots, Q_M) \in (P_n(\mathcal{X}))^M$ and for any $Q_m$, we use $T_m^n$ to denote the set of sequences $x = (x_1^n, \ldots, x_M^n)$ such that $x_i^n \in T_{Q_m}^n$ for all $i \in [M]$. Given any test $\phi_n$, the sample space $\mathcal{X}^{Mn}$ is separated into $(M + 1)$ disjoint regions: $\{\mathcal{A}_i(\phi_n)\}_{i \in [M]}$ and $\mathcal{A}_r(\phi_n)$ where

$$\mathcal{A}_i(\phi_n) = \{x \in \mathcal{X}^{Mn} : \phi_n(x) = H_i\},$$  

(198)

$$\mathcal{A}_r(\phi_n) = \left(\bigcup_{i \in [M]} \mathcal{A}_i\right)^c.$$  

(199)

We can then construct a type-based test as follows. Given any $\kappa$, for any $Q \in (P_n(\mathcal{X}))^M$,

- $\phi_n^+(Q) = H_i$ if at least $\kappa$ fractions of the sequences in the type class $T_Q^n$ are contained in the reject region, i.e.,

$$|T_Q^n \cap \mathcal{A}_i(\phi_n)| \geq \kappa|T_Q^n|. $$  

(200)
\( \phi_n^T (Q) = H_i \) if i) less than \( \kappa \) fractions of the sequences in the type class \( T_Q^n \) are contained in the reject region and ii) for all \( j \in [M] \), \( A_i (\phi_n) \) contains the most number of the sequences in the type class \( T_Q^n \), i.e.,

\[
|T_Q^n \cap A_i (\phi_n)| < \kappa |T_Q^n|, \quad \text{and} \quad |T_Q^n \cap A_i (\phi_n)| \geq \max_{j \in M_i} |T_Q^n \cap A_j (\phi_n)|.
\]

(201)

For any pair of distributions \((P_N, P_A)\), we can then relate the performances of an arbitrary test \( \phi_n \) and the constructed type-based test \( \phi_n^T \) as follows:

\[
\beta_i (\phi_n | P_T, P_A) = P_i \left\{ \bigcup_{j \in M_i} A_j (\phi_n) \right\}
\]

(202)

\[
= \sum_{j \in M_i} P (A_j (\phi_n))
\]

(203)

\[
= \sum_{j \in M_i} \sum_{Q \in (P_n (X))^M} P_i \{ A_j (\phi_n) \cap T_Q^n \}
\]

(204)

\[
\geq \sum_{j \in M_i} \sum_{Q \in (P_n (X))^M} \sum_{|T_Q^n \cap A_j (\phi_n)| < \kappa} P_i \{ A_j (\phi_n) \cap T_Q^n \}
\]

(205)

\[
\geq \sum_{j \in M_i} \sum_{Q \in (P_n (X))^M : \phi_n^T (Q) = H_j} \frac{1 - \kappa}{M - 1} P_i \{ T_Q^n \}
\]

(206)

\[
= \frac{1 - \kappa}{M - 1} \beta_i (\phi_n^T | P_T, P_A),
\]

(207)

and

\[
\zeta_i (\phi_n | P_T, P_A) = P_i \left\{ A_i (\phi_n) \right\}
\]

(209)

\[
= \sum_{Q \in (P_n (X))^M} P_i \{ A_i (\phi_n) \cap T_Q^n \}
\]

(210)

\[
\geq \sum_{Q \in (P_n (X))^M : |T_Q^n \cap A_i (\phi_n)| \geq \kappa} P_i \{ T_Q^n \}
\]

(211)

\[
\geq \kappa \sum_{Q \in (P_n (X))^M : |T_Q^n \cap A_i (\phi_n)| \geq \kappa} P_i \{ T_Q^n \}
\]

(212)

\[
= \kappa \zeta_i (\phi_n^T | P_T, P_A).
\]

(213)

G. Proof of Lemma 2

To prove Lemma 2, it suffices to prove that for any type-based test satisfying (119), if a tuple of types \( Q \) satisfies that

\[
g(Q) + \eta_{n,M} < \lambda,
\]

(214)

then we must have \( \phi_n^T (Q) = H_i \).

We will prove our claim by contradiction. Suppose our claim is not true. Then there exist types \( \bar{Q} = (\bar{Q}_1, \ldots, \bar{Q}_M) \in (P_n (X))^M \) such that for some \( i \in [M] \),

\[
\phi_n^T (\bar{Q}) = H_i \quad \text{and} \quad g(\bar{Q}) + \eta_{n,M} < \lambda.
\]

(215)

Note that (215) implies that there exists \((j, k) \in M^2\) such that \( j \neq k \) and

\[
G_j (\bar{Q}) + \eta_{n,M} < \lambda \quad \text{and} \quad G_k (\bar{Q}) + \eta_{n,M} < \lambda.
\]

(216)

Furthermore, either \( j \neq i \) or \( k \neq i \). Without loss of generality, we assume that \( j \neq i \).

Then, we have that for all pairs of distributions \((P_N, P_A)\), the misclassification error probability under hypothesis \( H_j \) can be lower bounded as follows:

\[
\beta_j (\phi_n^T | P_N, \bar{P}_A) \geq \sum_{Q \in (P_n (X))^M : \phi_n^T (Q) = H_i} P_j (T_Q^n)
\]

(217)

\[
\geq P_j (T_Q^n)
\]

(218)

\[
\geq (n + 1)^{-M_n} \exp \left( -n (D(\bar{Q}_j || \bar{P}_A) + \sum_{i \in M_j} D(\bar{Q}_i || \bar{P}_N)) \right).
\]

(219)
Now if we let $\tilde{P}_N = \hat{Q}_j$ and $\tilde{P}_N = \frac{\sum_{t \in M_j} \hat{Q}_t}{M-1}$, then
\begin{align}
\beta_j(\phi_n^T | \tilde{P}_N, \tilde{P}_N) & \geq (n + 1)^{-M} \exp(-nG_i(\tilde{Q})) \\
& = \exp\left(-n(G_i(\tilde{Q}) + \eta_{n,M})\right) \\
& > \exp(-n\lambda),
\end{align}
which contradicts that (119) holds. Therefore, we have show that for any type-based test $\phi_n^T$ satisfying (119), we must have $\phi_n^T(\tilde{Q}) = H_1$ for any $Q$ satisfying (214).

H. Justification of Properties of Exponent Tradeoff

We first prove that $\text{LD}_i(\lambda, P_N, P_A) = 0$ if and only if $\lambda \geq \text{GD}_M(P_N, P_A)$. Recall the definition of $\text{LD}_i(\cdot)$ in (9) and the definition of $G_i(\cdot)$ in (1). Note that for each $i \in [M]$, any $\lambda \in \mathbb{R}_+$ and any $(P_N, P_A)$, $\text{LD}_i(\lambda, P_N, P_A) = 0$ if there exists $(j, k) \in [M]^2$ such that $j \neq k$, $G_j(Q^*) \leq \lambda$ and $G_k(Q^*) \leq \lambda$ where $Q^*$ is a collection of distributions with $Q^*_i = P_A$ and $Q^*_i = P_N$ for all $t \in M_i$. For any $j \in M_i$, we have
\begin{align}
G_j(Q^*) &= \sum_{t \in (M_i \cap M_j)} D\left(Q_t^* \left\| \frac{\sum_{l \in M_j} Q_l}{M-1} \right\| \right) + D\left(Q_t^* \left\| \frac{\sum_{l \in M_j} Q_l}{M-1} \right\| \right) \\
& = (M - 2)D\left(P_N \left\| \frac{(M - 2)P_N + P_A}{M-1} \right\| \right) + D\left(P_A \left\| \frac{(M - 2)P_N + P_A}{M-1} \right\| \right)
\end{align}
Furthermore, if $j = i$, then
\begin{align}
G_j(Q^*) &= \sum_{t \in M_i} D\left(Q_t^* \left\| \frac{\sum_{l \in M_j} Q_l}{M-1} \right\| \right) = 0.
\end{align}
Combining (225) and (226), we have for each $i \in [M],$
\begin{align}
\max_{(j,k) \in [M]^2, j \neq k} \max\{G_j(Q^*), G_k(Q^*)\} &= \text{GD}_M(P_N, P_A).
\end{align}
Therefore, if $\lambda = \text{GD}_M(P_N, P_A)$, for each $i \in [M]$, we can find $(j, k) \in [M]^2$ such that $j \neq k$, $\max\{G_j(Q^*), G_k(Q^*)\} \leq \lambda$ and thus $\text{LD}_i(\lambda, P_N, P_A) = 0$. The justification is completed by the above argument with the fact that $\text{LD}_i(\lambda, P_N, P_A)$ is non-increasing in $\lambda$ for each $i \in [M]$ and any $(P_N, P_A)$.

We then prove (35). Since $\text{LD}_i(\lambda, P_N, P_A)$ is non-increasing in $\lambda$ for each $i \in [M]$ and any $(P_N, P_A)$, then we have
\begin{align}
sup_{\lambda \in \mathbb{R}_+} \text{LD}_i(\lambda, P_N, P_A) \\
\leq \text{LD}_i(0, P_N, P_A) \\
= \min_{(j,k) \in [M]^2, j \neq k} \min_{Q \in (P(X))^{M_i}, G_j(Q) = 0, G_k(Q) = 0} \left(D(Q_t^* \left\| P_A \right\| + \sum_{t \in M_i} D(Q_t^* \left\| P_A \right\|)\right) \\
= \min_{Q \in (P(X))^{M_i}} \left(D(Q \left\| P_A \right\| + (M - 1)D(Q \left\| P_N \right\|)\right),
\end{align}
where (230) follows from the definition of $G_i(\cdot)$ in (1). The proof of (35) is thus thus completed.

I. Justification of (111)

For any $i \in [M], j \in M_i, k \in M_{i,j}$, given any pair of distributions $(P_N, P_A)$,
\begin{align}
\text{Cov}_{P_{i,j}}[\xi_{i,j}(X_t | P_N, P_A)\xi_{i,k}(X_t | P_N, P_A)] \\
= E_{P_{i,j}}[\xi_{i,j}(X_t | P_N, P_A)\xi_{i,k}(X_t | P_N, P_A)] - E_{P_{i,j}}[\xi_{i,j}(X_t | P_N, P_A)]E_{P_{i,k}}[\xi_{i,k}(X_t | P_N, P_A)] \\
= E_{P_{i,j}}[\xi_{i,j}(X_t | P_N, P_A)\xi_{i,k}(X_t | P_N, P_A)] - (\text{GD}_M(P_N, P_A))^2,
\end{align}
where (233) follows from (109). The first term in (233) can be further calculated as follows:

$$\mathbb{E}_{P_i}[t_{i,j}(X_i|P_N, P_A) t_{i,k}(X_i|P_N, P_A)]$$

$$= \mathbb{E}_{P_i}\left[(t_1(X_{i,t}|P_N, P_A) + \sum_{l \in M_{i,j}} t_2(X_{l,t}|P_N, P_A))(t_1(X_{i,t}|P_N, P_A) + \sum_{l \in M_{i,k}} t_2(X_{l,t}|P_N, P_A))\right]$$

$$= \mathbb{E}_{P_i}\left[(t_1(X_{i,t}|P_N, P_A))^2 + \sum_{l \in M_{i,j}} t_1(X_{i,t}|P_N, P_A)t_2(X_{l,t}|P_N, P_A)\right]$$

$$+ \mathbb{E}_{P_i}\left[\sum_{l \in M_{i,j}} t_1(X_{i,t}|P_N, P_A)t_2(X_{l,t}|P_N, P_A)\right]$$

$$+ \mathbb{E}_{P_i}\left[\left(\sum_{l \in M_{i,j}} t_2(X_{l,t}|P_N, P_A)\right)^2\right].$$

(235)

We can calculate each term in (235). The first term in (235) satisfies

$$\mathbb{E}_{P_i}\left[(t_1(X_{i,t}|P_N, P_A))^2\right] = \mathbb{E}_{P_A}\left[(t_1(X|P_N, P_A))^2\right].$$

(236)

The second term in (235) satisfies

$$\mathbb{E}_{P_i}\left[\sum_{l \in M_{i,j}} t_1(X_{i,t}|P_N, P_A)t_2(X_{l,t}|P_N, P_A)\right] = \sum_{l \in M_{i,j}} \mathbb{E}_{P_i}[t_1(X_{i,t}|P_N, P_A)]\mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]$$

$$= (M-2)\mathbb{E}_{P_A}[t_1(X|P_N, P_A)]\mathbb{E}_{P_N}[t_2(X|P_N, P_A)].$$

(237)

Similarly, the third term in (235) satisfies

$$\mathbb{E}_{P_i}\left[\sum_{l \in M_{i,j}} t_1(X_{k,t}|P_N, P_A)t_2(X_{l,t}|P_N, P_A)\right] = (M-2)\mathbb{E}_{P_A}[t_1(X|P_N, P_A)]\mathbb{E}_{P_N}[t_2(X|P_N, P_A)].$$

(239)

Finally, the last term in (235) satisfies

$$\mathbb{E}_{P_i}\left[\left(\sum_{l \in M_{i,j}} t_2(X_{l,t}|P_N, P_A)\right)^2\right] + \mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]$$

$$+ \sum_{l \in M_{i,j}} \mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]\mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]$$

$$= \sum_{l \in M_{i,j}} \mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]\mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]$$

$$+ \sum_{l \in M_{i,j}} \mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]\mathbb{E}_{P_i}[t_2(X_{l,t}|P_N, P_A)]$$

$$= (M-2)(\mathbb{E}_{P_N}[t_2(X|P_N, P_A)])^2 + (M-3)\mathbb{E}_{P_N}\left[(t_2(X|P_N, P_A))^2\right]$$

$$+ (M-3)^2\mathbb{E}_{P_N}[t_2(X|P_N, P_A)]^2$$

$$= (M^2 - 5M + 7)(\mathbb{E}_{P_N}[t_2(X|P_N, P_A)])^2 + (M-3)\mathbb{E}_{P_N}\left[(t_2(X|P_N, P_A))^2\right].$$

(242)

Combining (233) to (244), we have that for any $i \in [M], j \in M_i, k \in M_{i,j},$

$$\text{Cov}_{P_i}[t_{i,j}(X_i|P_N, P_A) t_{i,k}(X_i|P_N, P_A)]$$

$$= -\text{GD}_{Mj}(P_N, P_A))^2 + \mathbb{E}_{P_A}\left[(t_1(X|P_N, P_A))^2\right] + 2(M-2)\mathbb{E}_{P_A}[t_1(X|P_N, P_A)]\mathbb{E}_{P_N}[t_2(X|P_N, P_A)]$$

$$+ (M^2 - 5M + 7)(\mathbb{E}_{P_N}[t_2(X|P_N, P_A)])^2 + (M-3)\mathbb{E}_{P_N}\left[(t_2(X|P_N, P_A))^2\right].$$

(245)
J. Justification of (195)

Given any $B \in S$, using the Berry-Esseen theorem and Taylor expansions in (175), (176), we have that for each $C \in S_B$

$$
P_B\{G_C(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}}) < G_B(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}})\}$$

$$
\leq P_B\{G_C(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}}) < G_B(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}}), X^n \in T_B(P_B)\} + P_B\{X^n \notin T_B(P_B)\}
$$

$$
\leq \mathbb{P}_B\left\{ \frac{1}{n} \sum_{t \in [n]} \left( \sum_{j \in (B\cap M_C)} t_{1,j,B}(j) X_{j,t} | B, C, P_N, P_B \right) + \sum_{j \in (M_B \cap M_C)} v_j X_{j,t} < O\left( \frac{\log n}{n}\right) \right\} + \frac{2M[X]}{n^2} \tag{246}
$$

$$
\leq Q\left( \frac{\sqrt{n}(GD(B, C, P_N, P_B) + O(\log n))}{\sqrt{V(B, C, P_N, P_B)}} \right) + \frac{6T(B, C, P_N, P_B)}{\sqrt{V(B, C, P_N, P_B)}} + \frac{2M[X]}{n^2} \tag{247}
$$

$$
\leq \exp\left( - \frac{n(GD(B, C, P_N, P_B) + O(\log n))^2}{2V(B, C, P_N, P_B)} \right) + \frac{6T(B, C, P_N, P_B)}{\sqrt{V(B, C, P_N, P_B)}} + \frac{2M[X]}{n^2} \tag{248}
$$

$$
=: \kappa_{T,n} = O\left( \frac{1}{\sqrt{n}} \right), \tag{249}
$$

where (247) follows from (173).

Using (250), we have that for any $B \in S$,

$$
P_B\{h_S(X^n) = \min_{C \in S_B} G_C(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}})\}$$

$$
= P_B\{G_B(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}}) \leq \min_{C \in S_B} G_C(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}})\} \tag{251}
$$

$$
\geq 1 - \sum_{C \in S_B} P_B\{G_C(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}}) < G_B(\hat{T}_{X_1}, \ldots, \hat{T}_{X_{n_B}})\} \tag{252}
$$

$$
\geq 1 - (|S| - 1) \kappa_{T,n} \tag{253}
$$

$$
= 1 - O\left( \frac{1}{\sqrt{n}} \right). \tag{254}
$$

Acknowledgement

The authors acknowledge three anonymous reviewers for many helpful comments and useful suggestions that help improve the quality of the current paper significantly.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM computing surveys (CSUR), vol. 41, no. 3, p. 15, 2009.
[2] Y. Li, S. Nitinawarat, and V. V. Veeravalli, “Universal outlier hypothesis testing,” IEEE Trans. Inf. Theory, vol. 60, no. 7, pp. 4066–4082, 2014.
[3] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” IEEE Trans. Inf. Theory, vol. 35, no. 2, pp. 401–408, 1989.
[4] L. Zhou, V. Y. F. Tan, and M. Motani, “Second-order optimal statistical classification,” Information and Inference: A Journal of the IMA, vol. 9, no. 1, pp. 81–111, 2020.
[5] Y. Bu, S. Zou, and V. V. Veeravalli, “Linear-complexity exponentially-consistent tests for universal outlying sequence detection,” IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2115–2128, 2019.
[6] S. Zou, Y. Liang, H. V. Poor, and X. Shi, “Nonparametric detection of anomalous data streams,” IEEE Transactions on Signal Processing, vol. 65, no. 21, pp. 5785–5797, Nov 2017.
[7] A. Tajer, V. V. Veeravalli, and H. V. Poor, “Outlying sequence detection in large data sets: A data-driven approach,” IEEE Signal Processing Magazine, vol. 31, no. 5, pp. 44–56, 2014.
[8] Y. Li, S. Nitinawarat, and V. V. Veeravalli, “Universal sequential outlier hypothesis testing,” Sequential Analysis, vol. 36, no. 3, pp. 309–344, 2017.
[9] S. Nitinawarat and V. V. Veeravalli, “Universal quickest outlier detection and isolation,” in IEEE ISIT, June 2015, pp. 770–774.
[10] K. Cohen and Q. Zhao, “Active hypothesis testing for anomaly detection,” IEEE Trans. Inf. Theory, vol. 61, no. 3, pp. 1452–1450, 2015.
[11] J. Unnikrishnan, “Asymptotically optimal matching of multiple sequences to source distributions and training sequences,” IEEE Trans. Inf. Theory, vol. 61, no. 1, pp. 452–468, 2015.
[12] H. He, L. Zhou, and V. V. Tan, “Distributed detection with empirically observed statistics,” IEEE Trans. Inf. Theory, vol. 65, no. 7, pp. 4349–4367, 2020.
[13] Y. F. F. Tan, “Asymptotic estimates in information theory with non-vanishing error probabilities,” Foundations and Trends® in Communications and Information Theory, vol. 11, no. 1–2, pp. 1–184, 2014.
[14] “Asymptotic estimates in information theory with non-vanishing error probabilities.” Foundations and Trends® in Communications and Information Theory, 2014, vol. 11, no. 1-2.
[15] E. L. Lehmann and J. P. Romano, Testing Statistical Hypotheses. Springer Science & Business Media, 2006.
[16] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2307–2359, 2010.
[17] H.-W. Hsu and I.-H. Wang, “On binary statistical classification from mismatched empirically observed statistics,” in IEEE ISIT. IEEE, 2020, pp. 2533–2538.
[18] T. M. Cover and J. A. Thomas, Elements of Information Theory. John Wiley & Sons, 2012.
[19] M. Tomamichel and V. Y. F. Tan, “Second-order coding rates for channels with state,” IEEE Trans. Inf. Theory, vol. 60, no. 8, pp. 4427–4448, 2014.
[20] V. Bentkus, “On the dependence of the Berry-Esseen bound on dimension,” J. Stat. Planning and Inference, vol. 113, pp. 385–402, 2003.
[21] A. C. Berry, “The accuracy of the Gaussian approximation to the sum of independent variates,” *Transactions of the American Mathematical Society*, vol. 49, no. 1, pp. 122–136, 1941.

[22] C.-G. Esseen, *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell, 1942.