Sequential Outlier Hypothesis Testing under Universality Constraints

Jun Diao and Lin Zhou
School of Cyber Science and Technology (CST), Beihang University, China
Emails: {jundiao,lzhou}@buaa.edu.cn

***Abstract*—We revisit sequential outlier hypothesis testing and derive bounds on the achievable exponents. Specifically, the task of outlier hypothesis testing is to identify the set of outliers that are generated from an anomalous distribution among all observed sequences where most are generated from a nominal distribution. In the sequential setting, one obtains a sample from each sequence per unit time until a reliable decision could be made. We assume that the number of outliers is known while both the nominal and anomalous distributions are *unknown*. For the case of exactly one outlier, our bounds on the achievable exponents are tight, providing exact large deviations characterization of sequential tests and strengthening a previous result of Li, Nitinawarat and Veeravalli (2017). In particular, we propose a sequential test that has bounded average sample size and better theoretical performance than the fixed-length test, which could not be guaranteed by the corresponding sequential test of Li, Nitinawarat and Veeravalli (2017). Our results are also generalized to the case of multiple outliers.**

***Index Terms*—Error Exponent, Large Deviations, Hypothesis testing, Anomaly Detection**

I. Introduction

Outlier hypothesis testing is a popular statistical inference problem [1]–[4], where one is asked to identify a set of outliers among a given number M of observed sequences. The majority of sequences are generated i.i.d. from a nominal distribution and the rest are generated i.i.d. from an anomalous distribution different from the nominal distribution. Both the nominal distribution and anomalous distributions are *unknown*.

The number of outliers can be assumed either known or unknown. When the number of outliers is known, the task is relatively simpler and corresponds to a generalization of classification [5], [6]. When the number of outliers is unknown, one could estimate the number of outliers and subsequently identify the set of outliers using a test for the known number case. As a compromise, one could also consider the case of at most T outliers, where an upper bound T on the number of outliers is known. When $T = 1$, the case is termed at most one [1], [3]. In this paper, for simplicity, we consider the case of known number of outliers and our results could be generalized to the case of unknown number of outliers by having an additional step to estimate the number of outliers.

Depending on the test design, a test could be fixed-length or sequential. When the sample size of each observed sequence is fixed, the corresponding test is a fixed-length test. When the sample size is a random variable depending on particular observations of sequences, the corresponding test is a sequential test. In a sequential test, one obtains a new sample from each sequence per unit time until one is confident to make a decision. The expected value of the sample size is also known as the expected stopping time. Since the generating distributions of sequences are unknown, for sequential tests, naturally, one can put a universal constraint either on the error probability or the average stopping time [7, Def. 2 and 3]. Specifically, for any pair of nominal and anomalous distributions, the error probability universality constraint requires the test to have the error probability bounded by a tolerable value $\beta \in (0, 1)$ under each hypothesis while the expected stopping time universality constraint requires that the expected stopping time under each hypothesis is bounded. Correspondingly, for fixed-length tests, only error probability universality constraint is valid since the sample size is fixed a-priori.

For both cases of at most one outlier and of at most T outliers, Li, Nitinawarat and Veeravalli [1] proposed generalized likelihood (GL) tests and proved the optimality of the test by having largest exponential decay rates of error probabilities when the number M of observed sequences tends to infinity. Subsequently, Li, Nitinawarat and Veeravalli [2] generalized the above results to the sequential setting under the error probability universality constraint. However, there are several limitations for the results in [2]. Firstly, only achievability results under the error probability universality constraint were derived. Without a matching converse result, the optimality of error exponents could not be guaranteed. Furthermore, the expected stopping time constraint was not considered, which leads to the undesired fact that the sequential tests might stop at very large sample sizes. Finally, it was only numerically shown that the sequential test outperforms the fixed-length test *only* when the average stopping time is relatively large [2, Figs. 1 and 2]. Without a theoretical guarantee, the benefit of the sequential design is not fully uncovered.

In this paper, for a slightly easier setting of exactly one outlier, we address all above limitations. Furthermore, we generalize our results to the case of multiple outliers when the number of outliers is known. Our main contribution is summarized in the following subsection.

A. Main Contributions

For the case of exactly one outlier, we refine the result in [2, Theorem 3.2] by deriving a matching converse result and re-

proving a simpler achievability part under the error probability universality constraint. Furthermore, we derive the exact error exponents under the expected stopping time constraint. In particular, in the achievability part, we propose a sequential test that has bounded average sample size under any pair of nominal and anomalous distributions and analytically show that the test could have strictly better performance than the fixed-length test in [1]. To compare the performance of the sequential tests under both universality constraints, we provide numerical examples to illustrate the achievable exponents and the expected stopping time of both tests, which imply that our proposed test under the expected stopping time universality constraint has much smaller average sample size and even better performance in certain scenarios. We also generalize our results to the case of multiple outliers and derive bounds on achievable exponents.

B. Other Related Works

We briefly recall other (non-exhausting) related works on outlier hypothesis testing. Bu, Zou and Veeravalli [4] proposed a low-complexity test for outlier hypothesis testing and showed that the test ensures exponential decay of error probabilities. Zhou, Wei and Hero [3] proposed an optimal threshold-based test under the generalized Neyman-Pearson criterion [5] and derived a second-order asymptotic approximation to the finite sample size performance. Zou *et al.* [8] used the maximum mean discrepancy metric to design a test for outlier hypothesis testing of continuous sequences and showed that the test is exponentially consistent

II. Problem Formulation and Existing Results

Notation

We use $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{N}$ to denote the set of real numbers, non-negative real numbers, and natural numbers respectively. Given any two integers $(a, b) \in \mathbb{N}^2$, we use $[a : b]$ to denote the set of integers $\{a, a+1, \ldots, b\}$ and use $[a]$ to denote $[1 : a]$. Random variables and their realizations are denoted by upper case variables (e.g., X) and lower case variables (e.g., x), respectively. All sets are denoted in calligraphic font (e.g., $\mathcal{X}$). Given any integer $N \in \mathbb{N}$, let $X^N := (X_1, \ldots X_N)$ be a random vector of length N and let $x^N = (x_1, \ldots, x_N)$ be a particular realization of X^N. The set of all probability distributions on a finite set $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$.

A. Problem Formulation

Consider a set of M observed sequences $\mathbf{X}^\tau := \left\{X_1^\tau, \ldots, X_M^\tau\right\}$, where τ is a random stopping time with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ and $\mathcal{F}_n$ is generated by σ-algebra $\sigma\{X_1, X_2, \ldots X_n\}$. Most sequences are generated i.i.d. from an unknown nominal distribution P_{N} while the rest of few sequences known as outliers are generated i.i.d. from an unknown anomalous distribution P_{A}. We first consider the case of exactly one outlier and then generalize the results to multiple outliers with the number of outliers known.

When there is exactly one outlier, the task is to design a test $\Phi = \{\tau, \phi_\tau\} : \mathcal{X}^{M\tau} \to \{\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_M\}$ that consists of a random stopping time τ and a decision rule ϕ_τ to classify among the following M hypotheses:

- $\mathrm{H}_i,\ i \in [M]$: the i-th sequence is the outlier.

To evaluate the performance of a test, we consider the misclassification error probability and the expected stopping time of a sequential test. Specifically, for each $i \in [M]$, the misclassification error probability is defined as follows:

$$\beta_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) := \mathbb{P}_i\{\Phi(\mathbf{X}^\tau) \neq \mathrm{H}_i\}, i \in [M], \tag{1}$$

where we define $\mathbb{P}_i(\cdot) := \Pr\{\cdot|\mathrm{H}_i\}$ to denote the joint distribution of observed sequences $\mathbf{X}^\tau$, where X_i^τ is generated i.i.d. from the anomalous distribution P_{A} and for each $j \in \mathcal{M}_i := \{j \in [M] : j \neq i\}$, X_j^τ is generated i.i.d. from the nominal distribution P_{N}. Furthermore, the expected stopping time under hypothesis H_i satisfies

$$\mathbb{E}_i[\tau] = \sum_{k=1}^{\infty} \mathbb{P}_i\{\tau > k\}. \tag{2}$$

Since there are two performance criteria, one could put a universal constraint on either one. Motivated by the analyses for sequential binary classification [7, Def. 2 and 3], we define the following two universality constraints on sequential tests.

Definition 1. *(Universality Constraint on the Error Probability): Given $\beta \in (0, 1)$ and a sequential test Φ, we say that Φ satisfies the universality constraint on the error probability with β if for any pair of distributions $(P_{\mathrm{N}}, P_{\mathrm{A}}) \in \mathcal{P}(\mathcal{X})^2$,*

$$\max_{i\in[M]} \beta_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq \beta. \tag{3}$$

For a sequential test satisfying the error probability universality constraint, we are interested in the following error exponent for each $i \in [M]$:

$$E_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) := \liminf_{n\to\infty} \frac{-\log\beta}{\mathbb{E}_i[\tau]}. \tag{4}$$

Definition 2. *(Universality Constraint on the Expected Stopping Time): Given $n \in \mathbb{N}$ and a sequential test Φ, we say that Φ satisfies the universality constraint on the expected stopping time with n if for any pair of distributions $(P_{\mathrm{N}}, P_{\mathrm{A}}) \in \mathcal{P}(\mathcal{X})^2$,*

$$\max_{i\in[M]} \mathbb{E}_i[\tau] \leq n. \tag{5}$$

For a sequential test satisfying the expected stopping time universality constraint, we are interested in the following error exponent for each $i \in [M]$:

$$E_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) := \liminf_{n\to\infty} \frac{-\log\beta_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}})}{n}. \tag{6}$$

B. Existing Results

To compare the performance of sequential tests and fixed-length test, we first recall the results of the fixed-length test Φ_{Li} by Li, Nitinawarat and Veeravalli [1]. To present the test, we need the following definition [3, Eq. (4)]. Given a tuple of distributions $\mathbf{Q} = (Q_1, \ldots, Q_M) \in \mathcal{P}(\mathcal{X})^M$, for each $i \in [M]$, define the following linear combination of

KL divergence terms between each single distribution and a mixture distribution:

$$\mathrm{G}_i(\mathbf{Q}) := \sum_{j\in\mathcal{M}_i} D\bigg(Q_j\bigg\|\frac{\sum_{l\in\mathcal{M}_i} Q_l}{M-1}\bigg), \tag{7}$$

where $\mathcal{M}_i = \{j \in [M] : j \neq i\}$. Note that $\mathrm{G}_i(\mathbf{Q})$ is used to measure the similarity of distributions $\mathbf{Q}$ except Q_i. The measure $\mathrm{G}_i(\mathbf{Q}) = 0$ if and only if $Q_j = Q$ for all $j \in \mathcal{M}_i$ for an arbitrary $Q \in \mathcal{P}(\mathcal{X})$.

For the case of exactly one outlier, the test in [1, Eq. (15)] applies the following minimal scoring function decision rule:

$$\Phi_{\mathrm{Li}}(\mathbf{x}^n) = \mathrm{H}_j, \text{ if } j = \arg\min_{i\in[M]} \mathrm{S}_i(\mathbf{x}^n), \tag{8}$$

where $\mathrm{S}_i(\mathbf{x}^n) = \mathrm{G}_i(\hat{T}_{x_1^n}, \ldots, \hat{T}_{x_M^n})$.

Li, Nitinawarat and Veeravalli derived the following result [1, Theorem 2].

Theorem 1. *Given any pair of distributions $(P_{\mathrm{N}}, P_{\mathrm{A}}) \in \mathcal{P}(\mathcal{X})^2$, the achievable error exponent of the fixed-length test satisfies that for each $i \in [M]$,*

$$E_i(\Phi_{\mathrm{Li}}|P_{\mathrm{A}}, P_{\mathrm{N}}) = \min_{\substack{\mathbf{Q}\in\mathcal{P}(\mathcal{X})^M:\\ \mathrm{G}_1(\mathbf{Q})\geq\mathrm{G}_2(\mathbf{Q})}} D(Q_1||P_{\mathrm{A}}) + \sum_{j\in[2,M]} D(Q_j||P_{\mathrm{N}}). \tag{9}$$

III. Main Results for Exactly One Outlier

We characterize the optimal error exponent of sequential outlier hypothesis testing under both the expected stopping time universality constraint and the error probability universality constraint. Specifically, for each case, we propose a corresponding sequential test using the empirical distributions of observed sequences and derive exact large deviations for error probabilities.

A. Error Probability Universality

1) Test Design and Intuition: Given any $\beta \in (0, 1)$ and $k \in \mathbb{N}$, define the set

$$\Psi_k(\mathbf{x}^k) := \Big\{l \in [M] : \ \mathrm{S}_l(\mathbf{x}^k) > g(\beta, k)\Big\}, \tag{10}$$

where the scoring function $\mathrm{S}_l(\mathbf{x}^k) = G_l(\hat{T}_{x_1^k}, \ldots, \hat{T}_{x_M^k})$ and the threshold satisfies

$$g(\beta, k) := \frac{-\log\big(\beta(|\mathcal{X}|-1)\big)}{k} + \frac{(M+1)|\mathcal{X}|\log(k+1)}{k}. \tag{11}$$

Under the error probability universality constraint, our sequential test $\Phi_{\mathrm{Ep}} = (\tau, \phi_\tau)$ consists of a random stopping time and the decision rule. The stopping time τ satisfies

$$\tau := \inf\big\{k \in \mathbb{N} : \ |\Psi_k(\mathbf{x}^k)| \geq M-1\big\}, \tag{12}$$

Note that for each $l \in [M]$, $\mathrm{S}_l(\mathbf{x}^k)$ measures the closeness of types of all sequences except the l-th sequence. Thus, sequential test Φ_{Ep} stops if the types of nominal samples and the outlier are far away for all $M-1$ possibilities of such mixtures. The threshold $g(\beta, k)$ determines how "far away" is measured, which increases with β and decreases with k.

At stopping time τ, our test uses the following decision rule:

$$\phi_\tau(\mathbf{x}^\tau) = \mathrm{H}_i, \text{ if } i = [M]\backslash\Psi_\tau(\mathbf{x}^\tau). \tag{13}$$

The above test generalizes the test for sequential classification in [9, Eq. (24)].

We now explain the intuitive reason why the above test works using the weak law of large numbers. Under hypothesis H_i, for each $j \in \mathcal{M}_i$, as the sample size k increases, the empirical distribution $\hat{T}_{x_j^k}$ of a nominal sequence x_j^k tends to the nominal distribution P_{N} while the empirical distribution $\hat{T}_{x_i^n}$ of the outlier tends to P_{A}. Thus, as k increases, the scoring function $\mathrm{S}_i(\mathbf{x}^k)$ tends to zero and scoring functions $\mathrm{S}_j(\mathbf{x}^k)$ for each $j \in \mathcal{M}_i$ tend to a positive real number. When k is sufficiently large, it follows from the weak law of large numbers that there exists $M-1$ scoring functions with positive values greater than the vanishing value of $g(\beta, k)$. Therefore, a correct decision could always be made asymptotically.

2) Main Results and Discussion: We need the following definition to present our results. Given any two distributions $(P, Q) \in \mathcal{P}(\mathcal{X})^2$ and any positive real number $\alpha \in \mathbb{R}_+$, the generalized Jensen-Shannon divergence [6, Eq. (2.3)] is defined as

$$\mathrm{GJS}(P, Q, \alpha) = \alpha D\left(P\Big\|\frac{\alpha P + Q}{1+\alpha}\right) + D\left(Q\Big\|\frac{\alpha P + Q}{1+\alpha}\right). \tag{14}$$

Theorem 2. *For any pair of distributions $(P_{\mathrm{N}}, P_{\mathrm{A}}) \in \mathcal{P}(\mathcal{X})^2$, our sequential test satisfies the error probability universality constraint with $\beta \in (0, 1)$ and the error exponent of our test satisfies that for each $i \in [M]$,*

$$E_i(\Phi_{\mathrm{Ep}}|P_{\mathrm{A}}, P_{\mathrm{N}}) \geq \mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2). \tag{15}$$

Conversely, for any sequential test Φ_β satisfying the error probability universality constraint with $\beta \in (0, 1)$, under any pair of distributions $(P_{\mathrm{A}}, P_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, the error exponent satisfies that for each $i \in [M]$,

$$E_i(\Phi_\beta|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq \mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2). \tag{16}$$

The proof of Theorem 3 is provided in Appendix A and B, which is inspired by the proof of [7, Theorem 2] for sequential binary classification under the error probability universality constraint. In the achievability proof, we show that for our sequential test, the random variable $\frac{\tau}{\log\beta}$ is uniformly integrable and subsequently we obtain the desired exponent by analyzing the convergence properties of $\frac{-\log\beta}{\mathbb{E}[\tau]}$. In the converse part, we use the binary KL divergence and apply the data processing inequality to upper bound error exponents.

Theorem 2 strengthens [2, Theorem 3.2] by deriving a matching converse result. We manage to do so for a slightly easier setting of exactly one outlier by excluding the null hypothesis where there might be no outliers. Furthermore, we propose another sequential test in addition to [2, Eq. (3.11)] and provide a relatively simpler achievability proof using our sequential test.

Since $\mathrm{GJS}(P,Q,\alpha)$ increases in α, it follows that as the number M of observed sequences increases, the achievable error exponent increases. This is consistent with our intuition because with more samples, our estimation of the nominal distribution is more accurate. Thus, it is easier to identify the outlier. In the extreme case of $M \to \infty$, the exponent equals to $D(P_{\mathrm{A}}\|P_{\mathrm{N}})$, which is exactly the performance of knowing the nominal distribution [2, Prop. 3.1].

B. Expected Stopping Time Universality

1) Test Design and Intuition: Under the expected stopping time universality constraint, our sequential test $\Phi_{\mathrm{Est}} = (\tau, \phi_\tau)$ consists of a random stopping time and a decision rule. The stopping time τ satisfies

$$\tau := \inf\{k \geq n-1 : \exists\ i \in [M] \text{ s.t. } \mathrm{S}_i(\mathbf{x}^k) \leq f(k)\}, \quad (17)$$

where the scoring function $\mathrm{S}_i(\mathbf{x}^k) = G_i(\hat{T}_{x_1^k}, \ldots, \hat{T}_{x_M^k})$ and the threshold is given by $f(k) = \frac{(M+1)|\mathcal{X}|\log(k+1)}{k}$.

Note that the sequential test Φ_{Est} stops if the types of all sequences except for outlying sequences are "close enough" to each other, where the threshold $f(k)$ is used to characterize the closeness level.

At the stopping time τ, using M observed sequences $\mathbf{x}_M^\tau$, our test applies the following minimal scoring function decision rule:

$$\phi_\tau(\mathbf{x}^\tau) = \mathrm{H}_i, \text{ if } i = i^*(\mathbf{x}^\tau), \quad (18)$$

where $i^*(\mathbf{x}^\tau)$ is the index of the scoring function with smallest value, i.e.,

$$i^*(\mathbf{x}^\tau) := \arg\min_{i\in[M]} \mathrm{S}_i(\mathbf{x}^\tau). \quad (19)$$

Our test generalizes the sequential classification test under expected stopping time universality in [7, Def. 7].

We now explain the intuitive reason why the above test works using the weak law of large numbers. As discussed below (13), under hypothesis H_i, the scoring function $\mathrm{S}_i(\mathbf{x}^k)$ tends to zero and scoring functions $\mathrm{S}_j(\mathbf{x}^k)$ for each $j \in \mathcal{M}_i$ tend to a positive real number as the sample size k increases. Therefore, when k is sufficiently large, for each $i \in [M]$, if i-th sequence is the outlier, our test stops and makes the correct decision H_i.

We remark that if the sample size is not large enough, the empirical distributions could be rather different from generating distributions, which might lead to decision error. To avoid such errors, similarly to [7, Def. 7], we set the minimal stopping time as $n-1$ for some integer $n \in \mathbb{N}$.

2) Main Results and Discussions: Given any pair of distributions $(P,Q) \in \mathcal{P}(\mathcal{X})^2$ and any $\alpha \in \mathbb{R}_+$, the Rényi Divergence of order α [10, Eq. (1)] is defined as

$$D_\alpha(P||Q) := \frac{1}{\alpha-1}\log\sum_{x\in\mathcal{X}} P(x)^\alpha Q(x)^{1-\alpha}. \quad (20)$$

The Rényi Divergence has the following variational form [7, Eq. (7)]:

$$D_{\frac{\alpha}{1+\alpha}}(P||Q) := \min_{V\in\mathcal{P}(\mathcal{X})} \alpha D(V||P) + D(V||Q). \quad (21)$$

Theorem 3. *Under any pair of distributions $(P_{\mathrm{N}}, P_{\mathrm{A}}) \in \mathcal{P}(\mathcal{X})^2$, our sequential test satisfies the expected stopping time universality constraint and the error exponent of our test satisfies: for each $i \in [M]$,*

$$E_i(\Phi_{\mathrm{Est}}|P_{\mathrm{A}}, P_{\mathrm{N}}) \geq D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}). \quad (22)$$

Conversely, for any sequential test Φ_n satisfying the expected stopping time universality constraint, under any pair of distributions $(P_{\mathrm{A}}, P_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, the error exponent satisfies that for each $i \in [M]$,

$$E_i(\Phi_n|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}). \quad (23)$$

The proof of Theorem 3 is provided in Appendix C and D. In the achievability part, we derive the lower bound of error exponents using method of types. In the converse part, we use the binary KL divergence and apply the data processing inequality to upper bound error exponents.

We make several remarks.

Firstly, Theorem 3 strengthens [2, Theorem 3.2] by considering sequential tests satisfying the expected stopping time universality constraint and deriving exact exponential decay rates of error probabilities. Compared with the error probability universality constraint in [2, Sec. 3.1.1] and Sec. III-A1, our proposed sequential test in this subsection has the property of having a bounded expected stopping time under any pair of nominal and anomalous distributions. This property is highly desired in practice since one desires the tests to stop early while the tests in [2, Sec. 3.1.1] and Sec. III-A1 could stop at very large sample sizes (cf. Section III-C for a numerical example).

Secondly, in contrast to lack of theoretical evidence that the sequential test of [2, Sec. 3.1.1] has better performance than the fixed-length test in (8), our sequential test has better theoretic performance than the fixed-length test in (8). This property is desired since the motivation of using sequential tests is to yield better performance. To clarify, we compare the exponents in Theorems 1 and 3. Given any distribution $Q \in \mathcal{P}(\mathcal{X})$, the distributions $\mathbf{Q} = (Q, Q, \ldots, Q, P_{\mathrm{N}})$ satisfy the constraints that $\mathrm{G}_1(\mathbf{Q}) \geq \mathrm{G}_2(\mathbf{Q})$, it follows from (9) that $E_i(\Phi_{\mathrm{Li}}|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq \min_{Q\in\mathcal{P}(\mathcal{X})} D(Q||P_{\mathrm{A}}) + (M-2)D(Q||P_{\mathrm{N}}) = D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}) = E_i(\Phi_{\mathrm{Est}}|P_{\mathrm{A}}, P_{\mathrm{N}})$. We numerically verify that $E_i(\Phi_{\mathrm{Est}}|P_{\mathrm{A}}, P_{\mathrm{N}})$ could be strictly greater than $E_i(\Phi_{\mathrm{Li}}|P_{\mathrm{A}}, P_{\mathrm{N}})$. Specifically, Under the nominal distribution $P_{\mathrm{N}} = [0.3, 0.7]$ and the anomalous distribution $P_{\mathrm{A}} = [0.1, 0.9]$, we have $E_i(\Phi_{\mathrm{Est}}|P_{\mathrm{A}}, P_{\mathrm{N}}) = 0.0934 > E_i(\Phi_{\mathrm{Li}}|P_{\mathrm{A}}, P_{\mathrm{N}}) = 0.0471$.

Finally, we compare the optimal error exponents in Theorems 2 and 3, i.e., $\mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2)$ and $D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}})$ respectively. It follows from [7, Remark 3] that for any $\alpha \in \mathbb{R}_+$ and any $(P_0, P_1) \in \mathcal{P}(\mathcal{X})^2$, $D_{\frac{\alpha}{1+\alpha}}(P_1||P_0) \geq \mathrm{GJS}(P_0, P_1, \alpha)$. Thus, setting $\alpha = M-2$ leads to $D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}) \geq \mathrm{GJS}(P_{\mathrm{A}}, P_{\mathrm{N}}, M-2)$, which implies that when $M = 3$, the sequential test under the expected stopping time constraint achieves better error exponent because $D_{\frac{1}{2}}(P_{\mathrm{N}}||P_{\mathrm{A}}) \geq \mathrm{GJS}(P_{\mathrm{A}}, P_{\mathrm{N}}, 1) = \mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, 1)$.

TABLE I
COMPARISON OF ACHIEVABLE ERROR EXPONENTS OF TWO SEQUENTIAL TESTS UNDER DIFFERENT UNIVERSALITY CONSTRAINTS

Parameters	$D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}\|\|P_{\mathrm{A}})$	$\mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2)$
$M=3$ $P_{\mathrm{N}}=[0.2,0.8]$ $P_{\mathrm{A}}=[0.4,0.6]$	**0.0493**	0.0483
$M=4$ $P_{\mathrm{N}}=[0.2,0.8]$ $P_{\mathrm{A}}=[0.4,0.6]$	0.0642	**0.0659**
$M=4$ $P_{\mathrm{N}}=[0.3,0.7]$ $P_{\mathrm{A}}=[0.1,0.9]$	**0.0939**	0.0830

However, since $\mathrm{GJS}(P_{\mathrm{A}}, P_{\mathrm{N}}, M-2) \neq \mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2)$ when $M > 3$, the performance comparison of sequential tests under two universality constraints depends on the nominal and anomalous distributions. To illustrate, in Table I, we calculate the exponents in Theorems 2 and 3 for various pairs of distributions.

C. Numerical Example

Consider the binary alphabet $\mathcal{X} = \{0, 1\}$ and set $M = 4$. We set the nominal distribution $P_{\mathrm{N}} = [0.25, 0.75]$ and the anomalous distribution $P_{\mathrm{A}} = [0.3, 0.7]$. We simulate the expected stopping times of our tests under two universality constraints using 5×10^4 independent experiments. The simulation shows that the test Φ_{Est} under the expected stopping time universality has expected stopping time of 6000 while the test Φ_{Ep} under the error probability universality test has expected stopping time of 20015. Thus, the numerical results verifies the advantage of sequential tests under the expected stopping time universality in terms of sample complexity.

IV. GENERALIZATION TO MULTIPLE OUTLIERS

Let $\mathcal{S}$ denote the set of all subsets of $[M]$ whose size is T, i.e., $\mathcal{S} := \{\mathcal{B} \subset [M] : |\mathcal{B}| = T\}$. Our task now is to design a test $\Phi = \{\tau, \phi_\tau\} : \mathcal{X}^{M\tau} \to \{\{\mathrm{H}_{\mathcal{B}}\}_{\mathcal{B}\in\mathcal{S}}\}$ with a stopping time τ and a corresponding decision rule ϕ_τ to classify among the following $|\mathcal{S}|$ hypotheses:

- $\mathrm{H}_{\mathcal{B}}$ where $\mathcal{B} \in \mathcal{S}$: the set of outlying sequences are sequences X_j^τ with $j \in \mathcal{B}$.

To evaluate the performance of a test, we use the following misclassification error exponent under each hypothesis $\mathrm{H}_{\mathcal{B}}$ with $\mathcal{B} \in \mathcal{S}$:

$$E_{\mathcal{B}}(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) := \liminf_{n\to\infty} \frac{-\log \mathbb{P}_{\mathcal{B}}\{\Phi(\mathbf{X}^\tau) \neq \mathrm{H}_{\mathcal{B}}\}}{n}. \quad (24)$$

To present our tests, we need the following definition [3, Eq. (42)]. Given a tuple of distributions $\mathbf{Q} = (Q_1, \ldots, Q_M) \in \mathcal{P}(\mathcal{X})^M$, for each $\mathcal{B} \in \mathcal{S}$, define

$$\mathrm{G}_{\mathcal{B}}(\mathbf{Q}) = \sum_{j\in\mathcal{M}_{\mathcal{B}}} D\Big(Q_j \Big\| \frac{\sum_{l\in\mathcal{M}_{\mathcal{B}}} Q_l}{M - |\mathcal{B}|}\Big), \quad (25)$$

where $\mathcal{M}_{\mathcal{B}} := [M]\backslash\mathcal{B} = \{i \in [M] : i \notin \mathcal{B}\}$. Note that $\mathrm{G}_{\mathcal{B}}(\mathbf{Q})$ measures the similarity of distributions $\mathbf{Q}$ except $\{Q_i\}_{i\in\mathcal{B}}$. The measure $\mathrm{G}_{\mathcal{B}}(\mathbf{Q}) = 0$ if and only if $Q_j = Q$ for all $j \in \mathcal{M}_{\mathcal{B}}$ for an arbitrary distribution $Q \in \mathcal{P}(\mathcal{X})$.

Under the expected stopping time universality constraint, our sequential test $\Phi_{\mathrm{Est}} = (\tau, \phi_\tau)$ consists of the random stopping time and the decision rule. The stopping time τ is defined as follows:

$$\tau := \inf\{k \geq n-1 : \exists\ \mathcal{B} \in \mathcal{S} \text{ s.t. } \mathrm{S}_{\mathcal{B}}(\mathbf{x}^k) \leq f(k)\}. \quad (26)$$

where the scoring function $\mathrm{S}_{\mathcal{B}}(\mathbf{x}^k) = \mathrm{G}_{\mathcal{B}}(\hat{T}_{x_1^k}, \ldots, \hat{T}_{x_M^k})$ and the threshold $f(k) = \frac{(M+1)|\mathcal{X}|\log(k+1)}{k}$. Our test applies the following minimal scoring function decision rule:

$$\phi_\tau(\mathbf{x}^\tau) = \arg\min_{\mathcal{B}\in\mathcal{S}} \mathrm{S}_{\mathcal{B}}(\mathbf{x}^\tau). \quad (27)$$

To present our results, we need the following two error exponent functions. Given any $\mathcal{B} \in \mathcal{S}$, for any nominal distribution P_{N} and anomalous distribution P_{A}, define

$$\mathrm{LD}_{\mathcal{B}}(P_{\mathrm{N}}, P_{\mathrm{A}}, M) := \min_{Q\in\mathcal{P}(\mathcal{X})} \min_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}} |\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}| D(Q||P_{\mathrm{A}}) + (M - |\mathcal{B}\cup\mathcal{C}|)D(Q||P_{\mathrm{N}}), \quad (28)$$

$$\tilde{\mathrm{LD}}_{\mathcal{B}}(P_{\mathrm{N}}, P_{\mathrm{A}}, M) := \min_{(Q_1,Q_2)\in\mathcal{P}(\mathcal{X})^2} \min_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}} |\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}| D(Q_1||P_{\mathrm{A}}) + (M - |\mathcal{B}\cup\mathcal{C}|)D(Q_1||P_{\mathrm{N}}) + |\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}| D(Q_2||P_{\mathrm{N}}) + |\mathcal{B}\cap\mathcal{C}| D(Q_2||P_{\mathrm{A}}). \quad (29)$$

Theorem 4. *Under any nominal distribution P_{N} and anomalous distribution P_{A}, our sequential test satisfies the expected stopping time universality constraint and the error exponent of our test satisfies that for each $\mathcal{B} \in \mathcal{S}$,*

$$E_{\mathcal{B}}(\Phi_{\mathrm{Est}}|P_{\mathrm{A}}, P_{\mathrm{N}}) \geq \mathrm{LD}_{\mathcal{B}}(P_{\mathrm{N}}, P_{\mathrm{A}}, M). \quad (30)$$

Conversely, for any sequential test Φ_n satisfying the expected stopping time universality constraint, under any pair of nominal distribution P_{N} and anomalous distribution P_{A}, the error exponent satisfies that for each $\mathcal{B} \in \mathcal{S}$,

$$E_{\mathcal{B}}(\Phi_n|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq \tilde{\mathrm{LD}}_{\mathcal{B}}(P_{\mathrm{N}}, P_{\mathrm{A}}, M). \quad (31)$$

The proof of Theorem 4 is similar to that of Theorem 3. The upper and lower bounds are not tight in general. However, when $T = 1$, the result in Theorem 4 specializes to Theorem 3.

V. CONCLUSION

We revisited sequential outlier hypothesis testing by proposing tests under two universality constraints and deriving bounds on the achievable error exponents. For the case of exactly one outlier, our results strengthen a previous result of [2] by having a matching converse result and analytically demonstrating the advantage of our sequential test. We also generalized our results to the case of multiple outliers when the number of outliers is known. In future, one can generalize our results to the case with an unknown number of outliers [3], [11], consider continuous observed sequences [8], [12] or propose low complexity tests [4], [13] that achieve performance close to the theoretical benchmarks derived in this paper.

REFERENCES

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

[2] ——, "Universal sequential outlier hypothesis testing," *Seq. Anal.*, vol. 36, no. 3, pp. 309–344, 2017.

[3] L. Zhou, Y. Wei, and A. O. Hero, "Second-order asymptotically optimal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 68, no. 6, pp. 3585–3607, 2022.

[4] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2115–2128, 2019.

[5] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401–408, 1989.

[6] L. Zhou, V. Y. F. Tan, and M. Motani, "Second-order asymptotically optimal statistical classification," *Information and Inference: A Journal of the IMA*, vol. 9, no. 1, pp. 81–111, 2020.

[7] C. Y. Hsu, C. F. Li, and I. H. Wang, "On universal sequential classification from sequentially observed empirical statistics," in *IEEE ITW*, 2022, pp. 642–647.

[8] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data streams," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5785–5797, 2017.

[9] M. Haghifam, V. Y. F. Tan, and A. Khisti, "Sequential classification with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 3095–3113, 2021.

[10] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[11] L. Zhou, Y. Wei, and A. Hero, "Asymptotics for outlier hypothesis testing," in *IEEE ISIT*, 2022, pp. 3303–3308.

[12] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Universal outlying sequence detection for continuous observations," in *IEEE ICASSP*, 2016, pp. 4254–4258.

[13] L. Xiong, B. Póczos, and J. Schneider, "Group anomaly detection using flexible genre models," *Advances in neural information processing systems*, vol. 24, 2011.

APPENDIX

A. Achievability Proof of Theorem 2

1) Error probability universality constraint: We first prove our sequential test in Sec. III-A1 satisfies the error probability universality constraint with β. Specifically, given any $\beta \in (0,1)$, for each $i \in [M]$,

$$
\begin{aligned}
&\beta_i(\phi_\tau|P_{\mathrm{A}}, P_{\mathrm{N}}) \\
&= \mathbb{P}_i\{\phi_\tau(\mathbf{X}^\tau) \neq \mathrm{H}_i\} && (32)\\
&\leq \sum_{k=1}^{\infty} \mathbb{P}_i\{\phi_\tau(\mathbf{X}^k) \neq \mathrm{H}_i\} && (33)\\
&\leq \sum_{k=1}^{\infty} \mathbb{P}_i\big\{\mathrm{S}_i(\mathbf{X}^k) > g(\beta,k)\big\} && (34)\\
&\leq \sum_{k=1}^{\infty} \sum_{\substack{\mathbf{Q}\in\mathcal{P}_k(\mathcal{X})^M:\\ \mathrm{G}_i(\mathbf{Q})>g(\beta,k)}} \exp\Big\{-k\Big(D(Q_i||P_{\mathrm{A}}) + \sum_{t\in\mathcal{M}_i} D(Q_t||P_{\mathrm{N}})\Big)\Big\} && (35)\\
&\leq \sum_{k=1}^{\infty} \sum_{\substack{\mathbf{Q}\in\mathcal{P}_k(\mathcal{X})^M:\\ \mathrm{G}_i(\mathbf{Q})>g(\beta,k)}} \exp\Big\{-k\Big(D(Q_i||P_{\mathrm{A}}) + (M-1)D\Big(\frac{\sum_{t\in\mathcal{M}_i} Q_t}{M-1}\Big\|P_{\mathrm{N}}\Big) + g(\beta,k)\Big)\Big\} && (36)\\
&\leq \sum_{k=1}^{\infty} (k+1)^{M|\mathcal{X}|} \exp\{-kg(\beta,k)\} && (37)\\
&= \sum_{k=1}^{\infty} (k+1)^{M|\mathcal{X}|} \beta(|\mathcal{X}|-1)(k+1)^{-(M+1)|\mathcal{X}|} && (38)\\
&\leq \beta(|\mathcal{X}|-1) \sum_{k=1}^{\infty} (k+1)^{-|\mathcal{X}|} && (39)\\
&\leq \beta(|\mathcal{X}|-1) \int_0^{\infty} (u+1)^{-|\mathcal{X}|}\mathrm{d}u && (40)\\
&= \beta(|\mathcal{X}|-1) \frac{1}{-|\mathcal{X}|+1}(u+1)^{-|\mathcal{X}|+1}\Big|_{u=0}^{u=\infty} && (41)\\
&= \beta, && (42)
\end{aligned}
$$

where (35) follows from the upper bound on the probability of the type class, (36) follows from $\mathrm{G}_i(\mathbf{Q}) > g(\beta,k)$ and

$$\sum_{j\in\mathcal{M}_i} D(Q_j||P_{\mathrm{N}}) = (M-1)D\Big(\frac{\sum_{t\in\mathcal{M}_i} Q_t}{M-1}\Big\|P_{\mathrm{N}}\Big) + \mathrm{G}_i(\mathbf{Q}), \tag{43}$$

(37) follows from the fact that the number of the set of types of length n satisfies $|\mathcal{P}^n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}$, (38) follows from the definition of $g(\beta,k)$ in Eq. (11), (40) follows from the similar manner as [7, Appendix C: 1)].

Then we have given any $\beta \in (0,1)$, for each $i \in [M]$, our sequential test satisfies $\beta_i(\phi_\tau|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq \beta$.

2) Achievable error exponents: Given $i \in [M]$, define the stopping time

$$\tau_i := \inf\big\{k \in \mathbb{N} : \forall j \in \mathcal{M}_i, \mathrm{S}_j(\mathbf{x}^k) > g(\beta,k)\big\}. \tag{44}$$

We first prove $-\frac{\tau_i}{\log\beta}$ converge to $\frac{1}{\mathrm{GJS}(P_{\mathrm{N}},P_{\mathrm{A}},M-2)}$ in probability.

Proof. Under hypothesis H_i, by the definition of the stopping time τ_i, we have

$$\min_{j\in\mathcal{M}_i} \mathrm{S}_j(\mathbf{x}^{\tau_i}) > g(\beta,\tau_i). \tag{45}$$

In the following, we show that $\tau_i \to \infty$ as $\beta \to 0$. Given $k \in \mathbb{N}$, we have

$$
\begin{aligned}
\mathbb{P}_i\{\tau_i \leq k\} &= \mathbb{P}_i\big\{\forall j \in \mathcal{M}_i, \mathrm{S}_j(\mathbf{x}^{\tau_i}) > g(\beta,\tau_i), \tau_i \leq k\big\} && (46)\\
&\leq \mathbb{P}_i\big\{kM\log(M-1) > -\log\big(\beta(|\mathcal{X}|-1)\big\} && (47)
\end{aligned}
$$

where (46) follows from Eq. (45) and (47) follows from $\forall j \in \mathcal{M}_i$, $\sum\limits_{t\in\mathcal{M}_j} D\Big(Q_t\Big\|\frac{\sum_{l\in\mathcal{M}_j} Q_l}{M-1}\Big) \le M\log(M-1)$ and $g(\beta,\tau_i) \ge g(\beta,k) \ge \frac{-\log(\beta(|\mathcal{X}|-1))}{k}$. Thus we can obtain that for each $i\in[M]$,

$$\mathbb{P}_i\{\tau_i \le k\} = 0, \quad \forall k < \frac{-\log\big(\beta(|\mathcal{X}|-1)\big)}{M\log(M-1)}. \tag{48}$$

Now we have $\tau_i \to \infty$ as $\beta \to 0$.

Furthermore, given $i\in[M]$, we obtain $\forall\ t\in\mathcal{M}_i$, $\hat{T}_{x_t^{\tau_i}} \to P_{\mathrm{N}}$ and $\hat{T}_{x_i^{\tau_i}} \to P_{\mathrm{A}}$ under hypothesis H_i when $\tau_i\to\infty$ as $\beta\to 0$ by the strong law of large numbers. Then with the continuity of KL-divergence, when $\tau_i\to\infty$, we have for each $j\in\mathcal{M}_i$,

$$\begin{aligned}\mathrm{S}_j(\mathbf{x}^{\tau_i}) &\to D\Big(P_{\mathrm{A}}\Big\|\frac{P_{\mathrm{A}}+(M-2)P_{\mathrm{N}}}{M-1}\Big) + (M-2)D\Big(P_{\mathrm{N}}\Big\|\frac{P_{\mathrm{A}}+(M-2)P_{\mathrm{N}}}{M-1}\Big)\\ &= \mathrm{GJS}(P_{\mathrm{N}},P_{\mathrm{A}},M-2).\end{aligned} \tag{49}$$

Recall the definition of the stopping time τ_i, we have

$$\min_{j\in\mathcal{M}_i} \mathrm{S}_j(\mathbf{x}^{\tau_i-1}) \le g(\beta,\tau_i-1). \tag{50}$$

When $\tau_i\to\infty$, both $g(\beta,\tau_i)$ and $g(\beta,\tau_i-1)\to -\frac{\log\beta}{\tau_i}$. Combining (50) and (49), we have

$$\lim_{\beta\to 0} -\frac{\log\beta}{\tau_i} \ge \mathrm{GJS}(P_{\mathrm{N}},P_{\mathrm{A}},M-2). \tag{51}$$

Then combining (45) and (49), we have

$$\lim_{\beta\to 0} -\frac{\log\beta}{\tau_i} \le \mathrm{GJS}(P_{\mathrm{N}},P_{\mathrm{A}},M-2). \tag{52}$$

Consequently, we conclude that

$$\lim_{\beta\to 0} -\frac{\tau_i}{\log\beta} = \frac{1}{\mathrm{GJS}(P_{\mathrm{N}},P_{\mathrm{A}},M-2)}. \tag{53}$$

□

To go from the convergence in probability to convergence in mean, it suffices to prove that the sequence of random variables $-\frac{\tau_i}{\log\beta}$ is uniformly integrable as $\beta\to 0$. To prove the uniformly integrable, we need the following lemma.

Lemma 5. *Given $k\in\mathbb{N},\ k\ge 1$, there exists $(n',c)\in\mathbb{R}_+^2$ such that $\mathbb{P}_i\{\tau_i\ge k\} \le \frac{1}{\beta}\exp\{-ck\}$ for any $k\ge n'$.*

Proof. Given $\varepsilon\in\mathbb{R}_+$ and $i\in[M]$, we have

$$\begin{aligned}&\mathbb{P}_i\{\tau_i\ge k\}\\ &\le \mathbb{P}_i\Big\{\exists\ j\in\mathcal{M}_i, \mathrm{S}_j(\mathbf{x}^{k-1}) \le g(\beta,k-1)\Big\} &(54)\\ &\le \sum_{j\in\mathcal{M}_i} \mathbb{P}_i\Big\{\mathrm{S}_j(\mathbf{x}^{k-1}) \le g(\beta,k-1)\Big\} &(55)\\ &\le \sum_{j\in\mathcal{M}_i} \mathbb{P}_i\Big\{\mathrm{S}_j(\mathbf{x}^{k-1}) \le g(\beta,k-1) \text{ and } D\big(\hat{T}_{x_i^{k-1}}\big\|P_{\mathrm{N}}\big)\le\varepsilon \text{ and } D\big(\hat{T}_{x_t^{k-1}}\big\|P_{\mathrm{N}}\big)\le\varepsilon, \forall\ t\in\mathcal{M}_i\Big\}\\ &\quad + \mathbb{P}_i\Big\{D\big(\hat{T}_{x_i^{k-1}}\big\|P_{\mathrm{N}}\big) > \varepsilon \text{ or } D\big(\hat{T}_{x_t^{k-1}}\big\|P_{\mathrm{N}}\big) > \varepsilon, \exists\ t\in\mathcal{M}_i\Big\}. &(56)\end{aligned}$$

The first term of (56) can be upper bounded as follows:

$$\begin{aligned}&\sum_{j\in\mathcal{M}_i} \mathbb{P}_i\Big\{\mathrm{S}_j(\mathbf{x}^{k-1}) \le g(\beta,k-1) \text{ and } D\big(\hat{T}_{x_i^{k-1}}\big\|P_{\mathrm{A}}\big)\le\varepsilon \text{ and } D\big(\hat{T}_{x_t^{k-1}}\big\|P_{\mathrm{N}}\big)\le\varepsilon, \forall\ t\in\mathcal{M}_i\Big\}\\ &\le \sum_{j\in\mathcal{M}_i} \mathbb{P}_i\Big\{\mathrm{S}_j(\mathbf{x}^{k-1}) \le g(\beta,k-1)+\mathrm{S}_i(\mathbf{x}^{k-1}) \text{ and } D\big(\hat{T}_{x_i^{k-1}}\big\|P_{\mathrm{A}}\big)\le\varepsilon \text{ and } D\big(\hat{T}_{x_t^{k-1}}\big\|P_{\mathrm{N}}\big)\le\varepsilon, \forall\ t\in\mathcal{M}_i\Big\} &(57)\\ &\le (M-1)\mathbb{P}_i\Big\{\alpha(P_{\mathrm{N}},P_{\mathrm{A}}) \le g(\beta,k-1)+\mathrm{S}_i(\mathbf{x}^{k-1})\Big\} &(58)\\ &\le \frac{M-1}{\beta(|\mathcal{X}|-1)} k^{(2M+1)|\mathcal{X}|}\exp\{-(k-1)\alpha(P_{\mathrm{N}},P_{\mathrm{A}})\}, &(59)\end{aligned}$$

where (57) follows from $\mathrm{S}_i(\mathbf{x}^k) \geq 0$, (58) follows from that when ε is sufficiently small, if $D(\hat{T}_{x_i^k}||P_{\mathrm{A}}) \leq \varepsilon$ and $D(\hat{T}_{x_t^k}||P_{\mathrm{N}}) \leq \varepsilon$ for all $t \in \mathcal{M}_i$, $\exists\ \alpha(P_{\mathrm{N}}, P_{\mathrm{A}}) > 0$ such that $\mathrm{S}_j(\mathbf{x}^k) > \alpha(P_{\mathrm{N}}, P_{\mathrm{A}})$ [2, Lemma B.1] and (59) follows from the upper bound on the probability of the type class and the number of the set of types. The second term of (56) can be upper bounded as follows using the upper bound of the probability of a type class:

$$\mathbb{P}_i\Big\{D\big(\hat{T}_{x_i^{k-1}}\big\|P_{\mathrm{A}}\big) > \varepsilon \text{ or } D\big(\hat{T}_{x_t^{k-1}}\big\|P_{\mathrm{N}}\big) > \varepsilon, \exists\ t \in \mathcal{M}_i\Big\} \leq Mk^{|\mathcal{X}|}\exp\{-(k-1)\varepsilon\}. \tag{60}$$

Thus combining (59) and (60), we have that for all $k \geq n'$ and for some c and $n' > 0$,

$$\mathbb{P}_i\{\tau_i \geq k\} \leq \frac{1}{\beta}\exp\{-ck\}. \tag{61}$$

□

Using Lemma 5 and [7, Lemma 5], we have that $\{\tau_i/\log\beta\}_{\beta\in(0,0.9]}$ is uniformly integrable. Therefore, we can obtain the convergence in mean of $\{\tau_i/\log\beta\}_{\beta\in(0,0.9]}$ from the convergence in probability in Eq. (53). Furthermore, we have $\tau \leq \tau_i$ by definition in (44). Then we have

$$\begin{aligned} E_i(\phi_\tau|P_{\mathrm{A}}, P_{\mathrm{N}}) &= \liminf_{\beta\to 0}\frac{-\log\beta}{\mathbb{E}_i[\tau]} && (62)\\ &\geq \liminf_{\beta\to 0}\frac{-\log\beta}{\mathbb{E}_i[\tau_i]} && (63)\\ &= \mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2). && (64)\end{aligned}$$

B. Converse Proof of Theorem 2

Given $(p, q) \in (0,1)^2$, define the binary KL-divergence as follows:

$$d(p,q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}. \tag{65}$$

The first-order derivative of $d(p,q)$ on q is:

$$\frac{\partial d(p,q)}{\partial q} = \frac{q-p}{q(1-q)}. \tag{66}$$

Thus, $d(p,q)$ is increasing in q when $q > p$ and decreasing in q when $p > q$.

Given $j \in [M]$ and any $i \in \mathcal{M}_j$, define the event $\mathcal{W} := \{\phi_\tau(\mathbf{x}^\tau) = i\}$. For any two pairs of distributions $(P_{\mathrm{A}}, P_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$ and $(\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, and any sequential test $\Phi = (\tau, \phi_\tau)$, we have

$$\begin{aligned} d\big(\mathbb{P}_i(\mathcal{W}), \tilde{\mathbb{P}}_j(\mathcal{W})\big) &\leq D(\mathbb{P}_i||\tilde{\mathbb{P}}_j)|_{\mathcal{F}_\tau} && (67)\\ &\leq \mathbb{E}_i\Bigg[\sum_{t\in[M]:t\neq i,j}\sum_{k=1}^{\tau}\log\frac{P_{\mathrm{N}}(X_{t,k})}{\tilde{P}_{\mathrm{N}}(X_{t,k})} + \sum_{k=1}^{\tau}\log\frac{P_{\mathrm{A}}(X_{i,k})}{\tilde{P}_{\mathrm{N}}(X_{i,k})} + \sum_{k=1}^{\tau}\log\frac{P_{\mathrm{N}}(X_{j,k})}{\tilde{P}_{\mathrm{A}}(X_{j,k})}\Bigg] && (68)\\ &\leq (M-2)\mathbb{E}_i[\tau]D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}}) + \mathbb{E}_i[\tau]D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}}) + \mathbb{E}_i[\tau]D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}), && (69)\end{aligned}$$

where (67) follows from data processing inequality of divergence and (69) follows from Doob's Optional Stopping Theorem.

Since $\beta_i(\phi_\tau|P_{\mathrm{A}}, P_{\mathrm{N}}) \to 0$ for each $i \in [M]$ as $n \to \infty$, we have

$$\begin{aligned} \mathbb{P}_i(\mathcal{W}) &= \mathbb{P}_i(\phi_\tau(\mathbf{x}^\tau) = i) = 1 - \beta_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) \to 1, && (70)\\ \tilde{\mathbb{P}}_j(\mathcal{W}) &= \tilde{\mathbb{P}}_j(\phi_\tau(\mathbf{x}^\tau) = i) \leq \tilde{\beta}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \to 0. && (71)\end{aligned}$$

Thus we obtain $\mathbb{P}_i(\mathcal{W}) > \tilde{\mathbb{P}}_j(\mathcal{W})$ and $1 - \beta_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) > \tilde{\beta}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}})$. Furthermore, we have

$$\begin{aligned} d\big(\mathbb{P}_i(\mathcal{W}), \tilde{\mathbb{P}}_j(\mathcal{W})\big) &\geq d\big(1-\beta_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}), \tilde{\beta}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}})\big) && (72)\\ &= -\log\tilde{\beta}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) && (73)\\ &\geq -\log\beta. && (74)\end{aligned}$$

where (72) follows from that $d(p,q)$ is decreasing in q when $p > q$, (73) follows from [7, Lemma 2] when $n \to \infty$ and (74) follows from the definition of error probability universality constraint.

Given $i \in [M]$, combining (69) and (74), we have

$$-\log\beta \leq (M-2)\mathbb{E}_i[\tau]D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}}) + \mathbb{E}_i[\tau]D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}}) + \mathbb{E}_i[\tau]D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}). \tag{75}$$

Therefore, for any sequential test $\Phi = (\tau, \phi_\tau)$ satisfying the expected stopping time universality constraint and any two pairs of distributions $(P_{\mathrm{A}}, P_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$ and $(\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, the type-$i$ error exponent satisfies

$$E_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq (M-2)D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}}) + D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}}) + D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}). \tag{76}$$

Since (76) is true for all $(\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, we can minimize the RHS of (76) with respect to $(\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}})$ and obtain

$$E_i(\Phi|P_{\mathrm{A}}, P_{\mathrm{N}}) \leq \min_{Q \in \mathcal{P}(\mathcal{X})} (M-2)D(P_{\mathrm{N}}||Q) + D(P_{\mathrm{A}}||Q) \tag{77}$$

$$= \mathrm{GJS}(P_{\mathrm{N}}, P_{\mathrm{A}}, M-2). \tag{78}$$

C. Achievability Proof of Theorem 3

1) Expected stopping time universality constraint: We first prove our sequential test in Sec. III-B1 satisfies the expected stopping time universality constraint. The average stopping time can be expressed as the following form: for each $i \in [M]$

$$\mathbb{E}_i[\tau] = \sum_{k=1}^{\infty} \mathbb{P}_i\{\tau > k\} = n - 1 + \sum_{k=n-1}^{\infty} \mathbb{P}_i\{\tau > k\}. \tag{79}$$

Then we upper bound the term $\sum_{k=n-1}^{\infty} \mathbb{P}_i\{\tau > k\}$ as follows,

$$\begin{aligned}
&\mathbb{P}_i\{\tau > k\} \\
&\leq \mathbb{P}_i\Big\{ \bigcap_{t=1}^{k} \mathrm{S}_i(\mathbf{X}^t) \geq f(t) \Big\} && (80)\\
&\leq \mathbb{P}_i\big\{ \mathrm{S}_i(\mathbf{X}^k) \geq f(k) \big\} && (81)\\
&= \sum_{\mathbf{x}^k \in \mathcal{X}^{Mk} : \mathrm{S}_i(\mathbf{x}^k) \geq f(k)} P_{\mathrm{A}}(x_i^k) \times \Big(\prod_{j \in \mathcal{M}_i} P_{\mathrm{N}}(x_j^k) \Big) && (82)\\
&= \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_i(\mathbf{Q}) \geq f(k)} P_{\mathrm{A}}(\mathcal{T}_{Q_i}^k) \times \Big(\prod_{j \in \mathcal{M}_i} P_{\mathrm{N}}(\mathcal{T}_{Q_j}^k) \Big) && (83)\\
&\leq \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_i(\mathbf{Q}) \geq f(k)} \exp\Big\{ -kD(Q_i||P_{\mathrm{A}}) - k \sum_{j \in \mathcal{M}_i} D(Q_j||P_{\mathrm{N}}) \Big\} && (84)\\
&\leq \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_i(\mathbf{Q}) \geq f(k)} \exp\Big\{ -k\Big(D(Q_i||P_{\mathrm{A}}) + (M-1)D\Big(\frac{\sum_{t \in \mathcal{M}_i} Q_t}{M-1} \Big\| P_{\mathrm{N}} \Big) + f(k) \Big) \Big\} && (85)\\
&\leq \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_i(\mathbf{Q}) \geq f(k)} \exp\{-kf(k)\} && (86)\\
&\leq (k+1)^{M|\mathcal{X}|} (k+1)^{-(M+1)|\mathcal{X}|} && (87)\\
&= (k+1)^{-|\mathcal{X}|}, && (88)
\end{aligned}$$

where (84) follows from the upper bound on the probability of the type class, (85) follows from the equation in Eq. (43) and $\mathrm{G}_i(\mathbf{Q}) \geq f(k)$ when $\tau > t$, (87) follows from the fact that the number of the set of types of length n satisfies $|\mathcal{P}^n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}$.

Thus for $n \geq 2$, we have

$$\sum_{k=n-1}^{\infty} \mathbb{P}_i\{\tau > k\} \leq \sum_{k=n-1}^{\infty} (k+1)^{-|\mathcal{X}|} \leq \int_{n-2}^{\infty} (u+1)^{-|\mathcal{X}|} \mathrm{d}u = \frac{1}{-|\mathcal{X}|+1} (u+1)^{-|\mathcal{X}|+1} \Big|_{u=n-2}^{u=\infty} \tag{89}$$

$$= \frac{(n-1)^{-(|\mathcal{X}|-1)}}{|\mathcal{X}|-1} \leq 1. \tag{90}$$

Combining with Eq. (79), we obtain that for each $i \in [M]$, under hypothesis H_i, $\mathbb{E}_i[\tau] \leq n$.

2) Achievable error exponents: Given any pair of distributions $(P_{\mathrm{N}}, P_{\mathrm{A}}) \in \mathcal{P}(\mathcal{X})^2$ and any $n \in \mathbb{N}, i \in [M]$, define the error exponent function

$$\Delta_i(n, P_{\mathrm{N}}, P_{\mathrm{A}}) := \min_{j \in \mathcal{M}_i} \min_{\substack{\mathbf{Q} \in \mathcal{P}_n(\mathcal{X})^M : \\ G_j(\mathbf{Q}) \leq f(n)}} D(Q_i||P_{\mathrm{A}}) + \sum_{t \in \mathcal{M}_i} D(Q_t||P_{\mathrm{N}}). \tag{91}$$

We upper bound the misclassification error probability as follows: under hypothesis H_i,

$$\begin{aligned}
&\beta_i(\phi_\tau|P_{\mathrm{A}},P_{\mathrm{N}})\\
&=\mathbb{P}_i\{\phi_\tau(\mathbf{X}^\tau)\neq i\} &(92)\\
&=\mathbb{P}_i\Big\{\bigcup_{j\in\mathcal{M}_i}\mathrm{S}_j(\mathbf{X}^\tau)\le f(\tau)\Big\} &(93)\\
&\le\mathbb{P}_i\Big\{\bigcup_{j\in\mathcal{M}_i}\bigcup_{k=n-1}^{\infty}\mathrm{S}_j(\mathbf{X}^k)\le f(k)\Big\} &(94)\\
&\le(M-1)\max_{j\in\mathcal{M}_i}\mathbb{P}_i\Big\{\bigcup_{k=n-1}^{\infty}\mathrm{S}_j(\mathbf{X}^k)\le f(k)\Big\} &(95)\\
&\le(M-1)\max_{j\in\mathcal{M}_i}\sum_{k=n-1}^{\infty}\mathbb{P}_i\{\mathrm{S}_j(\mathbf{X}^k)\le f(k)\} &(96)\\
&\le(M-1)\max_{j\in\mathcal{M}_i}\sum_{k=n-1}^{\infty}\sum_{\mathbf{Q}\in\mathcal{P}_k(\mathcal{X})^M:G_j(\mathbf{Q})\le f(k)}\exp\Big\{-k\Big(D(Q_i||P_{\mathrm{A}})+\sum_{t\in\mathcal{M}_i}D(Q_t||P_{\mathrm{N}})\Big)\Big\} &(97)\\
&\le(M-1)\sum_{k=n-1}^{\infty}\exp\Big\{-k\Big(\Delta_i(k,P_{\mathrm{N}},P_{\mathrm{A}})-\frac{M|\mathcal{X}|\log(k+1)}{k}\Big)\Big\} &(98)\\
&\le(M-1)\sum_{k=n-1}^{\infty}\exp\Big\{-k\Big(\Delta_i(n-1,P_{\mathrm{N}},P_{\mathrm{A}})-\frac{M|\mathcal{X}|\log n}{n-1}\Big)\Big\}, &(99)
\end{aligned}$$

where (97) follows from the same argument as Eq. (84) and (99) follows from both the function $\Delta_i(n,P_{\mathrm{N}},P_{\mathrm{A}})$ is increasing and $\frac{M|\mathcal{X}|\log n}{n-1}$ is decreasing in n.

As $f(n-1)\to 0$, $\frac{M|\mathcal{X}|\log n}{n-1}$ tends to 0 and $\Delta_i(n-1,P_{\mathrm{N}},P_{\mathrm{A}})$ tends to

$$\begin{aligned}
&\min_{(Q_1,Q_2)\in\mathcal{P}(\mathcal{X})^2}D(Q_1||P_{\mathrm{A}})+(M-2)D(Q_1||P_{\mathrm{N}})+D(Q_2||P_{\mathrm{N}})\\
&=\min_{Q\in\mathcal{P}(\mathcal{X})}D(Q||P_{\mathrm{A}})+(M-2)D(Q||P_{\mathrm{N}}) &(100)\\
&=D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}), &(101)
\end{aligned}$$

where the last equality follows from the variational form of the Rényi Divergence in (21). Then (99) can be upper bounded as

$$(M-1)\sum_{k=n-1}^{\infty}\exp\Big\{-kD_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}})\Big\}=(M-1)\frac{\exp\Big\{-(n-1)D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}})\Big\}}{1-\exp\Big\{-D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}})\Big\}}. \tag{102}$$

Finally, we obtain the error exponent of misclassification error probability satisfies: for each $i\in[M]$,

$$\begin{aligned}
E_i(\phi_\tau|P_{\mathrm{A}},P_{\mathrm{N}})&\ge\liminf_{n\to\infty}\Big\{\frac{n-1}{n}D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}})-\frac{\log(M-1)}{n}+\frac{1}{n}\log\Big(1-\exp\big\{D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}})\big\}\Big)\Big\} &(103)\\
&=D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}). &(104)
\end{aligned}$$

D. Converse Proof of Theorem 3

Recall the definition of the binary KL-divergence in Eq. (65) and its increasing and decreasing in Appendix B.

Given $j\in[M]$ and any $i\in\mathcal{M}_j$, define the event $\mathcal{W}:=\{\phi_\tau(\mathbf{x}^\tau)=i\}$. For any two pairs of distributions $(P_{\mathrm{A}},P_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^2$ and $(\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^2$, and any sequential test $\Phi=(\tau,\phi_\tau)$, we have

$$\begin{aligned}
d\big(\mathbb{P}_i(\mathcal{W}),\tilde{\mathbb{P}}_j(\mathcal{W})\big)&\le(M-2)\mathbb{E}_i[\tau]D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}})+\mathbb{E}_i[\tau]D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}})+\mathbb{E}_i[\tau]D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}) &(105)\\
&\le(M-2)nD(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}})+nD(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}})+nD(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}), &(106)
\end{aligned}$$

where (105) follows from the similar argument as Eq. (69) and (106) follows from the definition of expected stopping time universality constraint.

Analogously to Eq. (73) in Appendix B, we have

$$d\big(\mathbb{P}_i(\mathcal{W}),\tilde{\mathbb{P}}_j(\mathcal{W})\big)\ge-\log\tilde{\beta}_j(\Phi|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}}). \tag{107}$$

Given $j \in [M]$, combining (106) and (107), we have

$$-\log \tilde{\beta}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \leq (M-2)nD(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}}) + nD(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}}) + nD(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}). \tag{108}$$

Therefore, for any sequential test $\Phi = (\tau, \phi_\tau)$ satisfying the expected stopping time universality constraint and any two pairs of distributions $(P_{\mathrm{A}}, P_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$ and $(\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, the type-$j$ error exponent satisfies

$$\tilde{E}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) \leq (M-2)D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}}) + D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}}) + D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}}). \tag{109}$$

Since (109) is true for all $(P_{\mathrm{A}}, P_{\mathrm{N}}) \in \mathcal{P}(\mathcal{X})^2$, we can minimize the RHS of (109) with respect to $(P_{\mathrm{A}}, P_{\mathrm{N}})$ and obtain

$$\begin{aligned}\tilde{E}_j(\Phi|\tilde{P}_{\mathrm{A}}, \tilde{P}_{\mathrm{N}}) &\leq \min_{Q \in \mathcal{P}(\mathcal{X})} (M-2)D(Q||\tilde{P}_{\mathrm{N}}) + D(Q||\tilde{P}_{\mathrm{A}}) &(110)\\ &= D_{\frac{M-2}{M-1}}(P_{\mathrm{N}}||P_{\mathrm{A}}). &(111)\end{aligned}$$

E. Achievability Proof of Theorem 4

1) Expected stopping time universality constraint: The average stopping time can be expressed as the following form: for each $\mathcal{B} \in \mathcal{S}$,

$$\mathbb{E}_{\mathcal{B}}[\tau] = \sum_{k=1}^{\infty} \mathbb{P}_{\mathcal{B}}\{\tau > k\} = n - 1 + \sum_{k=n-1}^{\infty} \mathbb{P}_{\mathcal{B}}\{\tau > k\}. \tag{112}$$

Analogously to Appendix C1, we have that for each $\mathcal{B} \in \mathcal{S}$,

$$\begin{aligned}&\mathbb{P}_{\mathcal{B}}\{\tau > k\}\\ &\leq \mathbb{P}_{\mathcal{B}}\big\{\mathrm{S}_i(\mathbf{X}^k) \geq f(k)\big\} &(113)\\ &= \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_{\mathcal{B}}(\mathbf{Q}) \geq f(k)} \Big(\prod_{i \in \mathcal{B}} P_{\mathrm{A}}(\mathcal{T}^k_{Q_i})\Big) \times \Big(\prod_{j \in \mathcal{M}_{\mathcal{B}}} P_{\mathrm{N}}(\mathcal{T}^k_{Q_j})\Big) &(114)\\ &\leq \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_{\mathcal{B}}(\mathbf{Q}) \geq f(k)} \exp\Big\{-k\sum_{i \in \mathcal{B}} D(Q_i||P_{\mathrm{A}}) - k\sum_{j \in \mathcal{M}_{\mathcal{B}}} D(Q_j||P_{\mathrm{N}})\Big\} &(115)\\ &\leq \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_{\mathcal{B}}(\mathbf{Q}) \geq f(k)} \exp\Big\{-k\Big(\sum_{i \in \mathcal{B}} D(Q_i||P_{\mathrm{A}}) + (M-T)D\Big(\frac{\sum_{t \in \mathcal{M}_{\mathcal{B}}} Q_t}{M-T}\Big\|P_{\mathrm{N}}\Big) + f(k)\Big)\Big\} &(116)\\ &\leq \sum_{\mathbf{Q} \in \mathcal{P}_k(\mathcal{X})^M : \mathrm{G}_{\mathcal{B}}(\mathbf{Q}) \geq f(k)} \exp\{-kf(k)\} &(117)\\ &= (k+1)^{-|\mathcal{X}|}, &(118)\end{aligned}$$

where (116) follows from

$$\sum_{j \in \mathcal{M}_{\mathcal{B}}} D(Q_j||P_{\mathrm{N}}) = (M-T)D\Big(\frac{\sum_{t \in \mathcal{M}_{\mathcal{B}}} Q_t}{M-T}\Big\|P_{\mathrm{N}}\Big) + \mathrm{G}_{\mathcal{B}}(\mathbf{Q}), \tag{119}$$

and $\mathrm{G}_{\mathcal{B}}(\mathbf{Q}) \geq f(k)$ when $\tau > t$. Then we obtain that

$$\sum_{k=n-1}^{\infty} \mathbb{P}_{\mathcal{B}}\{\tau > k\} \leq \sum_{k=n-1}^{\infty} (k+1)^{-|\mathcal{X}|} \leq 1, \tag{120}$$

and thus

$$\mathbb{E}_{\mathcal{B}}[\tau] = n - 1 + \sum_{k=n-1}^{\infty} \mathbb{P}_{\mathcal{B}}\{\tau > k\} \leq n. \tag{121}$$

2) Achievable error exponents: We need the following definitions to present our proof. Recall the definition of $\mathcal{M}_{\mathcal{B}} = [M]\backslash\mathcal{B}$. Given any nominal distribution P_{N} and anomalous distributions $P_{\mathcal{B}}$, define

$$\Delta_{\mathcal{B}}(n, P_{\mathrm{N}}, P_{\mathrm{A}}) := \min_{\mathcal{C} \in \mathrm{S}_{\mathcal{B}}} \min_{\substack{\mathbf{Q} \in \mathcal{P}_n(\mathcal{X})^M : \\ \mathrm{G}_{\mathcal{C}}(\mathbf{Q}) \leq f(n)}} \sum_{i \in \mathcal{B}} D(Q_i||P_{\mathrm{A}}) + \sum_{t \in \mathcal{M}_{\mathcal{B}}} D(Q_t||P_{\mathrm{N}}). \tag{122}$$

Define $\mathrm{S}_{\mathcal{B}} := \{\mathcal{C} \in \mathcal{S} : \mathcal{C} \neq \mathcal{B}\}$ for any $\mathcal{B} \in \mathcal{S}$.

Analogously to Appendix C2, we upper bound the misclassification error probability under hypothesis $\mathrm{H}_{\mathcal{B}}$ as follows,

$$\begin{aligned}
&\beta_{\mathcal{B}}(\phi_\tau|P_{\mathrm{A}},P_{\mathrm{N}})\\
&=\mathbb{P}_{\mathcal{B}}\Big\{\bigcup_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}}\mathrm{S}_{\mathcal{C}}(\mathbf{X}^\tau)\le f(\tau)\Big\} &(123)\\
&\le\mathbb{P}_{\mathcal{B}}\Big\{\bigcup_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}}\bigcup_{k=n-1}^{\infty}\mathrm{S}_{\mathcal{C}}(\mathbf{X}^k)\le f(k)\Big\} &(124)\\
&\le(|S|-1)\max_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}}\sum_{k=n-1}^{\infty}\mathbb{P}_{\mathcal{B}}\{\mathrm{S}_{\mathcal{C}}(\mathbf{X}^k)\le f(k)\} &(125)\\
&\le(|S|-1)\max_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}}\sum_{k=n-1}^{\infty}\sum_{\mathbf{Q}\in\mathcal{P}_k(\mathcal{X})^M:\mathrm{G}_{\mathcal{C}}(\mathbf{Q})\le f(k)}\exp\Big\{-k\Big(\sum_{i\in\mathcal{B}}D(Q_i||P_{\mathrm{A}})+\sum_{t\in\mathcal{M}_{\mathcal{B}}}D(Q_t||P_{\mathrm{N}})\Big)\Big\} &(126)\\
&\le(|S|-1)\sum_{k=n-1}^{\infty}\exp\Big\{-k\Big(\Delta_{\mathcal{B}}(k,P_{\mathrm{N}},P_{\mathrm{A}})-\frac{M|\mathcal{X}|\log(k+1)}{k}\Big)\Big\} &(127)\\
&\le(|S|-1)\sum_{k=n-1}^{\infty}\exp\Big\{-k\Big(\Delta_{\mathcal{B}}(n-1,P_{\mathrm{N}},P_{\mathrm{A}})-\frac{M|\mathcal{X}|\log n}{n-1}\Big)\Big\}. &(128)
\end{aligned}$$

As $f(n-1)\to 0$, $\frac{M|\mathcal{X}|\log n}{n-1}\to 0$ and $\Delta_{\mathcal{B}}(n-1,P_{\mathrm{N}},P_{\mathrm{A}})\to$

$$\begin{aligned}
&\min_{(Q,\{Q_j\}_{j\in\mathcal{C}})\in\mathcal{P}(\mathcal{X})^{T+1}}\min_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}}|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(Q||P_{\mathrm{A}})+(M-|\mathcal{B}\cup\mathcal{C}|)D(Q||P_{\mathrm{N}})+\sum_{l\in\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}}D(Q_l||P_{\mathrm{N}})+\sum_{t\in\mathcal{B}\cap\mathcal{C}}D(Q_t||P_{\mathrm{A}})\\
&=\min_{Q\in\mathcal{P}(\mathcal{X})}\min_{\mathcal{C}\in\mathrm{S}_{\mathcal{B}}}|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(Q||P_{\mathrm{A}})+(M-|\mathcal{B}\cup\mathcal{C}|)D(Q||P_{\mathrm{N}}) &(129)\\
&:=\mathrm{LD}_{\mathcal{B}}(P_{\mathrm{N}},P_{\mathrm{A}},M). &(130)
\end{aligned}$$

Then (128) can be upper bounded as follows:

$$(|S|-1)\sum_{k=n-1}^{\infty}\exp\{-k\cdot\mathrm{LD}(P_{\mathrm{N}},P_{\mathrm{A}},M)\}=(|S|-1)\frac{\exp\{-(n-1)\mathrm{LD}(P_{\mathrm{N}},P_{\mathrm{A}},M)\}}{1-\exp\{-\mathrm{LD}(P_{\mathrm{N}},P_{\mathrm{A}},M)\}}. \tag{131}$$

Thus we obtain that the achievable error exponent of misclassification error probability satisfies: for each $\mathcal{B}\in\mathcal{S}$,

$$E_{\mathcal{B}}(\phi_\tau|P_{\mathcal{B}},P_{\mathrm{N}})\ge\mathrm{LD}_{\mathcal{B}}(P_{\mathrm{N}},P_{\mathrm{A}},M). \tag{132}$$

F. Converse Proof of Theorem 4

We have stated that the function $d(p,q)$ in Eq. (65) is increasing in q when $q>p$ and decreasing in q when $p>q$ in Appendix D. Given $\mathcal{B}\in\mathcal{S}$, define the event $\mathcal{W}:=\{\phi_\tau(\mathbf{x}^\tau)=\mathcal{B}\}$. Given $\mathcal{B}\in\mathcal{S}$ and $\mathcal{C}\in\mathrm{S}_{\mathcal{B}}$, for any two pairs of distributions $(P_{\mathrm{A}},P_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^2$ and $(\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^2$, and any sequential test $\Phi=(\tau,\phi_\tau)$, we have

$$\begin{aligned}
&d\big(\mathbb{P}_{\mathcal{B}}(\mathcal{W}),\tilde{\mathbb{P}}_{\mathcal{C}}(\mathcal{W})\big)\\
&\le D(\mathbb{P}_{\mathcal{B}}||\tilde{\mathbb{P}}_{\mathcal{C}})|_{\mathcal{F}_\tau} &(133)\\
&\le\mathbb{E}_{\mathcal{B}}\Big[\sum_{t\in\mathcal{M}_{\mathcal{B}\cup\mathcal{C}}}\sum_{k=1}^{\tau}\log\frac{P_{\mathrm{N}}(X_{t,k})}{\tilde{P}_{\mathrm{N}}(X_{t,k})}+\sum_{i\in\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}}\sum_{k=1}^{\tau}\log\frac{P_{\mathrm{A}}(X_{i,k})}{\tilde{P}_{\mathrm{N}}(X_{i,k})}+\sum_{j\in\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}}\sum_{k=1}^{\tau}\log\frac{P_{\mathrm{N}}(X_{j,k})}{\tilde{P}_{\mathrm{A}}(X_{j,k})}+\sum_{l\in\mathcal{B}\cap\mathcal{C}}\sum_{k=1}^{\tau}\log\frac{P_{\mathrm{A}}(X_{l,k})}{\tilde{P}_{\mathrm{A}}(X_{l,k})}\Big] &(134)\\
&\le\mathbb{E}_{\mathcal{B}}[\tau]\Big((M-|\mathcal{B}\cup\mathcal{C}|)D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}})+|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}})+|\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}|D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}})+|\mathcal{B}\cap\mathcal{C}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{A}})\Big) &(135)\\
&\le n\Big((M-|\mathcal{B}\cup\mathcal{C}|)D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}})+|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}})+|\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}|D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}})+|\mathcal{B}\cap\mathcal{C}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{A}})\Big), &(136)
\end{aligned}$$

where (133) follows from data processing inequality of divergence, (135) follows from Doob's Optional Stopping Theorem and (136) follows from the definition of expected stopping time universality constraint.

Since $\beta_{\mathcal{B}}(\Phi|P_{\mathrm{A}},P_{\mathrm{N}})\to 0$ for each $i\in[M]$ as $n\to\infty$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{B}}(\mathcal{W})=\mathbb{P}_{\mathcal{B}}(\phi_\tau(\mathbf{x}^\tau)=\mathcal{B})=1-\beta_{\mathcal{B}}(\Phi|P_{\mathrm{A}},P_{\mathrm{N}})\to 1, &(137)\\
&\tilde{\mathbb{P}}_{\mathcal{C}}(\mathcal{W})=\tilde{\mathbb{P}}_{\mathcal{C}}(\phi_\tau(\mathbf{x}^\tau)=\mathcal{B})\le\tilde{\beta}_{\mathcal{C}}(\Phi|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\to 0. &(138)
\end{aligned}$$

Thus we obtain $\mathbb{P}_{\mathcal{B}}(\mathcal{W}) > \tilde{\mathbb{P}}_{\mathcal{C}}(\mathcal{W})$ and $1-\beta_{\mathcal{B}}(\Phi|P_{\mathrm{A}},P_{\mathrm{N}}) > \tilde{\beta}_{\mathcal{C}}(\Phi|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})$. Furthermore, we have

$$d\big(\mathbb{P}_{\mathcal{B}}(\mathcal{W}),\tilde{\mathbb{P}}_{\mathcal{C}}(\mathcal{W})\big) \geq d\big(1-\beta_{\mathcal{B}}(\Phi|P_{\mathrm{A}},P_{\mathrm{N}}),\tilde{\beta}_{\mathcal{C}}(\Phi|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\big) \tag{139}$$

$$= -\log\tilde{\beta}_{\mathcal{C}}(\Phi|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}}), \tag{140}$$

where (139) follows from that $d(p,q)$ is decreasing in q when $p>q$ and (140) follows from [7, Lemma 2] when $n\to\infty$.

Given $\mathcal{C}\in\mathcal{S}$, combining (136) and (140), we have

$$\begin{aligned}&-\log\tilde{\beta}_{\mathcal{C}}(\Phi|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\\ &\leq n\min_{\mathcal{B}\in\mathrm{S}_{\mathcal{C}}}(M-|\mathcal{B}\cup\mathcal{C}|)D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}})+|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}})+|\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}|D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}})+|\mathcal{B}\cap\mathcal{C}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{A}}).\end{aligned} \tag{141}$$

Therefore, for any sequential test $\Phi=(\tau,\phi_\tau)$ satisfying the expected stopping time universality constraint and any two pairs of distributions $(P_{\mathrm{A}},P_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^2$ and $(\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^2$, the error exponent satisfies: given $\mathcal{C}\in\mathcal{S}$

$$\begin{aligned}&\tilde{E}_{\mathcal{C}}(\phi_\tau|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}})\\ &\leq\min_{\mathcal{B}\in\mathrm{S}_{\mathcal{C}}}(M-|\mathcal{B}\cup\mathcal{C}|)D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{N}})+|\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}|D(P_{\mathrm{N}}||\tilde{P}_{\mathrm{A}})+|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{N}})+|\mathcal{B}\cap\mathcal{C}|D(P_{\mathrm{A}}||\tilde{P}_{\mathrm{A}}).\end{aligned} \tag{142}$$

Since (142) is true for all $(P_{\mathrm{A}},P_{\mathrm{N}})\in\mathcal{P}(\mathcal{X})^{T+1}$, we can minimize the RHS of (142) with respect to $(P_{\mathrm{A}},P_{\mathrm{N}})$ and obtain

$$\begin{aligned}\tilde{E}_{\mathcal{C}}(\phi_\tau|\tilde{P}_{\mathrm{A}},\tilde{P}_{\mathrm{N}}) \leq \min_{(Q_1,Q_2)\in\mathcal{P}(\mathcal{X})^2}\min_{\mathcal{B}\in\mathrm{S}_{\mathcal{C}}}&(M-|\mathcal{B}\cup\mathcal{C}|)D(Q_1||\tilde{P}_{\mathrm{N}})+|\mathcal{C}\cap\mathcal{M}_{\mathcal{B}}|D(Q_1||\tilde{P}_{\mathrm{A}})\\ &+|\mathcal{B}\cap\mathcal{M}_{\mathcal{C}}|D(Q_2||\tilde{P}_{\mathrm{N}})+|\mathcal{B}\cap\mathcal{C}|D(Q_2||\tilde{P}_{\mathrm{A}})\end{aligned} \tag{143}$$

$$= \tilde{\mathrm{LD}}_{\mathcal{C}}(\tilde{P}_{\mathrm{N}},\tilde{P}_{\mathrm{A}},M). \tag{144}$$