Sequential Classification With Empirically Observed Statistics

Mahdi Haghifam, *Member, IEEE*, Vincent Y. F. Tan, *Senior Member, IEEE*, and Ashish Khisti, *Member, IEEE*

***Abstract*— Motivated by real-world machine learning applications, we consider a statistical classification task in a sequential setting where test samples arrive sequentially. In addition, the generating distributions are unknown and only a set of empirically sampled sequences are available to a decision maker. The decision maker is tasked to classify a test sequence which is known to be generated according to either one of the distributions. In particular, for the binary case, the decision maker wishes to perform the classification task with minimum number of the test samples, so, at each step, she declares that either hypothesis 1 is true, hypothesis 2 is true, or she requests for an additional test sample. We propose a classifier and analyze the type-I and type-II error probabilities. We demonstrate the significant advantage of our sequential scheme compared to an existing non-sequential classifier proposed by Gutman. Finally, we extend our setup and results to the multi-class classification scenario and again demonstrate that the variable-length nature of the problem affords significant advantages as one can achieve the same set of exponents as Gutman's fixed-length setting but without having the rejection option.**

***Index Terms*— Sequential classification, empirically sampled sequences, error exponents, variable-length.**

I. Introduction

QUICK and accurate classification is crucial in many real-life applications. For instance, to diagnose haematologic diseases based on blood test results, a physician wishes to detect the pattern, deviations, and relations in the blood samples of a patient as quickly as possible to make treatment plans. Similar challenges can be found in a broad range of applications such as genomics analysis, finance, and abnormal detection where there is an inherent trade-off between speed and accuracy.

In many real-world applications, classical *hypothesis testing* is infeasible due to the fact that the probability distributions of the sources are unknown. In practice, one often encounters *classification* problems in which one has access to training samples and is required to *classify* a set of test samples according to which distribution this set is generated from. To incorporate the real-life requirement of classifying the test samples as quickly as possible, one can consider the *sequential statistical classification* setup. This setup addresses the problem of classifying test samples given training samples with the additional requirement that the decision maker is required to make his/her decision based on as *few* tests samples as possible; it is however, known that all the test samples originate from the same distribution.

The problem of classification using empirically observed statistics has been studied in many prior works. Gutman in [2] formulated a problem in which a decision maker has access to two training sequences which are generated according to two distinct and unknown distributions. Then, a fixed-length test sequence is given to the decision maker, and the decision maker is tasked to classify the test sequence. For this problem, Gutman proposes an asymptotically optimal test. The results in [2] are obtained in the asymptotic regime when the length of the training sequences tends to infinity. In this regard, the non-asymptotic and second-order performance of the Gutman's test is analyzed in [3] where it is shown that Gutman's test is, in fact, second-order optimal. Moreover, Ziv [4] studied the relationship between test rules for the binary classification problem and universal data compression methods. Unnikrishnan and Naini [5] and Unnikrishnan [6] extended Gutman's proposed test for the case with multiple test sequences and obtain an optimal test rule for a certain matching task between multiple test sequences. Furthermore, Unnikrishnan and Huang in [7] showed how one can apply the results on the weak convergence of the test statistic to obtain better approximations for the error probabilities for statistical classification in the finite sample size setting. Kelly *et al.* [8] considered the classification problem with empirically observed statistics for large alphabet sources. They consider a scenario in which the alphabet size grows with the length of the training and the test sequences, and the authors characterized the maximum growth rate of the alphabet size for which consistent classification is possible. The related problem of closeness testing has been investigated in [9], [10]. Another related problem in this area is estimating properties of distributions using empirically observed statistics. This problem has been considered in various setups such as estimation of the support of distribution [11], [12] and the estimation of the order of a finite-state Markov chain [13], etc.

Manuscript received December 2, 2019; revised October 22, 2020; accepted February 5, 2021. Date of publication February 12, 2021; date of current version April 21, 2021. This article was presented in part at the 2019 IEEE Information Theory Workshop (ITW). *(Corresponding author: Mahdi Haghifam.)*

Mahdi Haghifam and Ashish Khisti are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: mahdi.haghifam@mail.utoronto.ca; akhisti@ece.utoronto.ca).

Vincent Y. F. Tan is with the Department of Electrical and Computer Engineering and the Department of Mathematics, National University of Singapore, Singapore 119077 (e-mail: vtan@nus.edu.sg).

Communicated by E. Gassiat, Associate Editor for Probability and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3059272.

Digital Object Identifier 10.1109/TIT.2021.3059272

Recently, Acharya *et al.* in [14] proposed an optimal method for the estimation of certain properties of distribution using empirically observed statistics which is applicable for a wide range of property estimation problems. The authors in [15] studied the problem of distributed detection in the setting that the central node has access to noisy test and training sequences. Finally, [16] considered the Gutman's setup with the difference is that there is a mismatch between the generating distribution of the test sequence and that of the training sequences, which they called "mismatch". In this setup an optimal classifier was proposed in [16].

In this paper, we consider an information-theoretic formulation of sequential classification. Recall that in the simple sequential binary hypothesis testing scenario, a decision maker is given a variable-length test sequence and knows that it is either generated in an i.i.d. fashion from one of the *known* distributions P_1 or P_2. It is well-known that the sequential probability ratio test (SPRT) is optimal for sequential binary hypothesis testing [17]. However, we consider a scenario that the decision maker *does not know both* generating distributions, i.e., P_1 and P_2. Instead the decision maker has access to two fixed-length training sequences, one is drawn i.i.d. from P_1 and the other i.i.d. from P_2. Then, the task of the decision maker is to classify a test sequence which is drawn i.i.d. from either P_1 or P_2. The decision maker observes the test sequence sequentially and may choose when to stop sampling once she is sufficiently confident. At that time, she makes a final decision. Also, we extend our framework beyond the binary classification setting and consider a sequential multi-class classification problem without the rejection option.

A. Main Contribution

Our contribution in the paper can be summarized as follows. In this paper, we extend the statistical classification problem with empirically observed statistics to the case when the decision maker observes the test sequence sequentially. First, we consider the binary classification problem and propose a test for the sequential setting. We analyse the performance of this test in terms of type-I and type-II error exponents (Theorem 2 and Corollary 1). Then, we show that this test outperforms Gutman's test [2] in terms of Bayesian error exponent (Theorem 3). Furthermore, we generalize the problem setup to the multi-class classification. For this case, we describe an achievable scheme and provide a characterization of its error exponents (Theorem 5 and Corollary 3). As a consequence of our results, we show that our test achieves the same performance as that of Gutman's but our test is arguably simpler as it does not consist of the rejection option (Theorem 6).

B. Paper Outline

The remainder of the paper is organized as follows. Section II describes the problem setup and summarizes the main results for the binary classification. In Section III we extend the problem to the multi-class classification problem and presents our main results. Sections IV-B and IV-D are devoted to the proofs of the results provided in Section II and III.

C. Notations

For each $m \in \mathbb{N}$, let $[m] \triangleq \{1, \dots, m\}$. The set of all discrete distributions on alphabet $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. We use upper and lower letters to denote random variables and their realizations, respectively. For a vector of length n, we use the notation $x^n = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$. Given a vector $x^n = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$, the type or empirical distribution is defined as

$$\widehat{Q}_{x^n}(a) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i = a\}, \quad \forall\, a \in \mathcal{X},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Also $\mathcal{T}_n$ represents the set of types with denominator n. The set of all sequences of length n with type Q is denoted by Γ_Q^n (we sometimes omit n if it is clear from the context). In addition, we use $\mathbb{E}[\cdot]$ to denote expectation, and, when not clear from context, we use a subscript to indicate the distribution with respect to which the expectation is being taken; e.g., $\mathbb{E}_Q[\cdot]$ denotes expectation with respect to the distribution Q. Other notation concerning the method of types follows [18, Chapter 11] and [19]. If P is a distribution on $\mathcal{X}$ then P^n is the n-fold i.i.d. product measure on $\mathcal{X}^n$, i.e.,

$$P^n(x^n) = \prod_{i=1}^{n} P(x_i), \quad \forall\, x^n \in \mathcal{X}^n.$$

The notion $a_n \doteq b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Similarly we can define $\dot{\leq}$ and $\dot{\geq}$. For other information-theoretic notations we use the standard definitions, see e.g., [18]. Also, for a function $f : \mathbb{N} \to \mathbb{R}$, we say that $g(n) = O(f(n))$ if $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$, $g(n) = o(f(n))$ if $\lim_{n\to\infty} |f(n)/g(n)| = 0$, and $g(n) = \omega(f(n))$ if $\liminf_{n\to\infty} |f(n)/g(n)| = \infty$. Hence, for example, $o(1)$ denotes a vanishing sequence.

II. Binary Sequential Classification

A. Problem Statement and Existing Results

We assume that a decision maker has two training sequences of length N. The first and second training sequences are generated in an i.i.d. manner according to $P_1 \in \mathcal{P}(\mathcal{X})$ and $P_2 \in \mathcal{P}(\mathcal{X})$ respectively. The underlying distributions (P_1, P_2) are *unknown* but *fixed* (i.e., remain unchanged throughout). The training sequences are denoted as $X_1^N \in \mathcal{X}^N$ and $X_2^N \in \mathcal{X}^N$. We fix a certain distribution P_{i^*} where $i^* \in \{1, 2\}$ which is unknown to the decision maker. Then, at each time $n \in \mathbb{N}$, a *test sample* $Y_n \in \mathcal{X}$ as generated from P_{i^*} and Y_n is given to the decision maker. The objective of the decision maker is to decide between the following two hypotheses:

- H_1: The test sequence up to the current time $\{Y_k\}_{k=1}^n$ (which is generated i.i.d. according to P_{i^*}) and the first training sequence X_1^N are generated according to the same distribution.

- H_2: The test sequence up to the current time $\{Y_k\}_{k=1}^n$ and the second training sequence X_2^N are generated according to the same distribution.

To achieve this goal, the decision maker at each time n can take three actions:

1) Stop drawing a new test sample and declare the test sequence and the first training sequence are generated according to the same distribution.
2) Stop drawing a new test sample and declare the test sequence and the second training sequence are generated according to the same distribution.
3) Continue to draw a new test sample from P_{i^*}.

In contrast to sequential hypothesis testing [20, Section 15.3] where the two distributions are known, in this setup, the decision maker does not know either of the distributions. Instead, the only information decision maker has about P_1 and P_2 is through the two training sequences X_1^N and X_2^N generated in an i.i.d. fashion according to P_1 and P_2 respectively. Moreover, the problem considered here is different from [2]–[6] where the classification is studied for the cases that the length of the test sequence is *fixed* prior to the decision making. In our setup, we let the length of the test sequence be *random*. In fact, the total number of samples is a *stopping time* determined by the decision maker's action, i.e., this is a *variable-length setting*. Next, we provide a precise formulation of this problem. We begin with the definition of the test for the aforementioned setup.

Definition 1 (Test): A *test* is a pair $\Phi = (T, d)$ where

- The integer-valued random variable $T \in \mathbb{N}$ is a stopping time with respect to the filtration $\mathcal{F}_n = \sigma\{X_1^N, X_2^N, Y_1, \ldots, Y_n\}$ generated by the training samples and the test samples up to time n.
- The map $d : \left(X_1^N, X_2^N, Y^T\right) \rightarrow \{H_1, H_2\}$ is a $\mathcal{F}_T$-measurable decision rule.

Definition 2 (Type-I and Type-II Error Probabilities): For a test $\Phi = (T, d)$, the *type-I* and *type-II error probabilities* are defined as

$$\mathrm{P}_i^{\mathrm{err}}(\Phi) = \mathrm{P}_i\left(d\left(X_1^N, X_2^N, Y^T\right) \neq H_i\right)$$

for $i \in \{1, 2\}$ respectively. Here P_i denote the probability distribution under H_i. Also, note that for any $S \subseteq \mathcal{X}^N \times \mathcal{X}^N \times \mathcal{X}^n$ we have $\mathrm{P}_i(S) = \sum_{(x_1^N, x_2^N, y^n)\in S} P_1^N(X_1^N)P_2^N(X_2^N)P_i^n(y^n)$. Also, $\mathbb{E}_i$ denotes the expectation under P_i.

Definition 3 (Error Exponents): For a test $\Phi = (T, d)$ such that $\mathrm{P}_i^{\mathrm{err}}(\Phi) \to 0$ as $N \to \infty$ and for $i \in \{1, 2\}$, we define the type-i error exponent as

$$\mathsf{e}_i(\Phi) = \liminf_{N\to\infty} \frac{-\log \mathrm{P}_i^{\mathrm{err}}(\Phi)}{\mathbb{E}_i[T]}$$

where $\mathbb{E}_i[T]$ represents the expected value of the stopping time under hypothesis H_i.

Remark 1: Note that the error event, i.e., $d\left(X_1^N, X_2^N, Y^T\right) \neq H_i$, and the random variable T depend on N. Furthermore, $\mathbb{E}_i[T]$ indicates the average number of the test samples under H_i before the decision is made.

Gutman [2] considers the setup in which the decision maker has a test sequence Y^n of *fixed length* n which is independently generated from X_1^N and X_2^N. Note as $N \to \infty$, n also diverges but we have $\lim_{N,n\to\infty} \frac{N}{n} = \alpha$. For example, we can think always think of $N = \lfloor n\alpha \rfloor$. To present Gutman's results, we need the following definition.

Definition 4 (Generalized Jensen-Shannon (GJS) Divergence): Given $\alpha \in \mathbb{R}_+$ and $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, the *generalized Jensen-Shannon (GJS) divergence* is defined as

$$\mathrm{GJS}(P_1, P_2, \alpha) = \alpha \mathrm{D}\big(P_1 \| P_\alpha\big) + \mathrm{D}\big(P_2 \| P_\alpha\big). \tag{1}$$

where $P_\alpha = \frac{\alpha P_1 + P_2}{1+\alpha}$.

Theorem 1 summarizes Gutman's main results concerning with achievable error exponents and the converse results for the binary classification task using non-adaptive tests.

Theorem 1: (Gutman [2, Thm. 1]) Let $\frac{N}{n} = \alpha$ and $\lambda \in \mathbb{R}_+$. Then Gutman's decision rule

$$\Phi_{\mathrm{GUT}}(\lambda, \alpha) = \begin{cases} H_1 & \text{if } \mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \alpha\right) \leq \lambda, \\ H_2 & \text{if } \mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \alpha\right) > \lambda, \end{cases} \tag{2}$$

has the following type-I and type-II error exponents

$$\mathsf{e}_1(\Phi_{\mathrm{GUT}}(\lambda, \alpha)) = \liminf_{N\to\infty} \frac{-\log \mathrm{P}_1^{\mathrm{err}}(\Phi_{\mathrm{GUT}}(\lambda, \alpha))}{N/\alpha} \geq \lambda, \tag{3}$$

$$\mathsf{e}_2(\Phi_{\mathrm{GUT}}(\lambda, \alpha)) = \liminf_{N\to\infty} \frac{-\log \mathrm{P}_2^{\mathrm{err}}(\Phi_{\mathrm{GUT}}(\lambda, \alpha))}{N/\alpha} > F(\alpha, \lambda), \tag{4}$$

where

$$\begin{aligned} F(\alpha, \lambda) \triangleq \min_{(Q_1, Q_2)\in\mathcal{P}(\mathcal{X})^2} \quad & \alpha \mathrm{D}\big(Q_1 \| P_1\big) + \mathrm{D}\big(Q_2 \| P_2\big) \\ \text{subject to} \quad & \mathrm{GJS}(Q_1, Q_2, \alpha) \leq \lambda. \end{aligned} \tag{5}$$

Also, Gutman's decision rule is optimal in the sense that among all non-adaptive decision rules Φ, satisfying $\mathsf{e}_1(\Phi) \geq \lambda$ for all pairs of distinct distributions $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, one has $\mathsf{e}_2(\Phi_{GUT}(\lambda, \alpha)) \geq \mathsf{e}_2(\Phi)$.

Remark 2: The intuition for Gutman's test in (2) and the bounds in (3) and (4) are as follows. The rule in (2) posits that we should choose H_1 if the type of the first set of training samples X_1^N and the test sequence Y^n are "close". The appropriate measure of closeness in this scenario is the GJS with parameter α because the GJS arises naturally as the exponent when one uses Sanov's theorem to establish the exponential rates of decay of the error probabilities and carefully takes into account the different lengths of the training and test sequences (see Lemma 2 and Lemma 5). The bounds in (3) and (4) are natural consequences in view of Sanov's theorem. In fact, similar to the Neyman-Pearson rule, Gutman's test is optimal (cf. [3]).

B. Main Results

We now describe our sequential classification test. Fix a threshold parameter $\gamma \in \mathbb{R}_+$. The proposed test for the sequential classification is $\Phi_{\mathrm{seq}}(\gamma) = (T_{\mathrm{seq}}, d_{\mathrm{seq}})$ where T_{seq}

and d_{seq} are defined as

$$T_{\text{seq}} = \inf \Bigg\{ n \geq 1 : \exists\, i \in \{1,2\} \text{ such that } \\ n\text{GJS}\left(\widehat{Q}_{X_i^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \geq \gamma N \Bigg\} \wedge N^2, \tag{6}$$

and

$$d_{\text{seq}} = \begin{cases} H_1 & \text{if } T_{\text{seq}}\text{GJS}\left(\widehat{Q}_{X_2^N}, \widehat{Q}_{Y^{T_{\text{seq}}}}, \frac{N}{T_{\text{seq}}}\right) \geq \gamma N \\ H_2 & \text{if } T_{\text{seq}}\text{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^{T_{\text{seq}}}}, \frac{N}{T_{\text{seq}}}\right) \geq \gamma N \end{cases}, \tag{7}$$

respectively. In (6), $\wedge$ denotes the pairwise minimum operation, i.e., $a \wedge b = \min(a, b)$.

We can view the the decision rule in (7) as assigning a *score* at each time $n \in \mathbb{N}$, i.e., $\text{score}_i[n] = n\text{GJS}\left(\widehat{Q}_{X_i^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right)$, to each class. At the first time that the score of one of the class exceeds the threshold, i.e., γN, the decision maker outputs the class with the least score. As an illustrative example, Figure 1 shows a realization of $\Phi_{\text{seq}}(\gamma)$ for two ternary source distributions $P_1 = [0.1, 0.7, 0.2]$ and $P_2 = [0.05, 0.55, 0.4]$. The threshold is $\gamma = 0.02$. The test sequence is drawn from P_2 and the length of the training sequence is $N = 400$. Note that the stopping time defined in (6) is $T_{\text{seq}} = 141$ since at $n = 141$, the score for class 1, i.e., $n\text{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right)$, exceeds the threshold γ. Therefore, based on (7), we declare H_2 as the final decision.

Remark 3: Note that in the definition of T_{seq} in (6), N^2 can be replaced by any function $h(\cdot) : \mathbb{N} \to \mathbb{N}$ with the following properties: 1) $h(N) = \omega\left(N^{\frac{3}{2}}\right)$ and, 2) $\frac{1}{N}\log h(N) = o(1)$. It can be verified that $h(N) = N^2$ satisfies the aforementioned conditions.

Before, presenting our main result on binary classification, we need the following definition.

Definition 5: The *Chernoff information* between two probability mass functions $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{X})$ is defined as

$$C(P, Q) \triangleq -\min_{\eta \in [0,1]} \log \sum_{x \in \mathcal{X}} P(x)^{\eta} Q(x)^{1-\eta}. \tag{8}$$

In the next theorem, we present the main result of this section which is on the properties of test $\Phi_{\text{seq}}(\gamma)$ and the achievable type-I and type-II error exponents of the proposed test for the binary classification problem. The proof of Theorem 2 is provided in Section IV-B.

Theorem 2: Fix pair $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$ and $\gamma \in (0, C(P_1, P_2)]$. Define $\beta_\gamma^\star \in \mathbb{R}_+$ to be the solution of

$$\text{GJS}\left(P_2, P_1, \beta_\gamma^\star\right) = \gamma \beta_\gamma^\star. \tag{9}$$

Similarly, define $\theta_\gamma^\star \in \mathbb{R}_+$ to be the solution of

$$\text{GJS}\left(P_1, P_2, \theta_\gamma^\star\right) = \gamma \theta_\gamma^\star. \tag{10}$$

Then, the proposed test has the following properties:

- $\text{P}_1^{\text{err}}\left(\Phi_{\text{seq}}(\gamma)\right) \stackrel{.}{\leq} \exp\left(-N\gamma\right)$
- $\mathbb{E}_1\left[T_{\text{seq}}\right] = \frac{N}{\beta_\gamma^\star}\left(1 + o(1)\right)$
- $\text{P}_2^{\text{err}}\left(\Phi_{\text{seq}}(\gamma)\right) \stackrel{.}{\leq} \exp\left(-N\gamma\right)$
- $\mathbb{E}_2\left[T_{\text{seq}}\right] = \frac{N}{\theta_\gamma^\star}\left(1 + o(1)\right)$

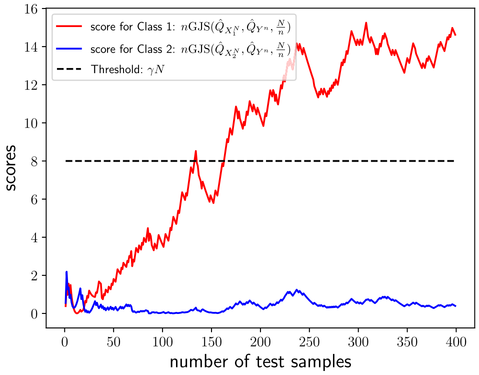

Fig. 1. A realization of the sequential test $\Phi_{\text{seq}}(\gamma) = (T_{\text{seq}}, d_{\text{seq}})$ in (6) and (7).

The following corollary summarizes our results on the achievable Type-I and Type-II error exponents.

Corollary 1: The achievable Type-I and Type-II error exponents of the proposed test are given by

$$\mathsf{e}_1\left(\Phi_{\text{seq}}(\gamma)\right) \geq \text{GJS}\left(P_2, P_1, \beta_\gamma^\star\right), \tag{11}$$

$$\mathsf{e}_2\left(\Phi_{\text{seq}}(\gamma)\right) \geq \text{GJS}\left(P_1, P_2, \theta_\gamma^\star\right), \tag{12}$$

where $\theta_\gamma^\star$ and $\beta_\gamma^\star$ are given by in (9) and (10) respectively.

Remark 4: In this remark we provide a *partial* converse bound for our results in Corollary 1. It is easy to observe that the performance of the SPRT provides an upper bound on the performance of our test. Therefore, combining [20, Thm. 15.3] with (11) and (12), we obtain the following upper bound

$$\mathsf{e}_1\left(\Phi_{\text{seq}}(\gamma)\right)\mathsf{e}_2\left(\Phi_{\text{seq}}(\gamma)\right) \leq \text{D}\big(P_1 \| P_2\big)\text{D}\big(P_2 \| P_1\big),$$

and lower bound

$$\mathsf{e}_1(\Phi_{\text{seq}}(\gamma))\mathsf{e}_2(\Phi_{\text{seq}}(\gamma)) \geq \text{GJS}\left(P_2, P_1, \beta_\gamma^\star\right)\text{GJS}\left(P_1, P_2, \theta_\gamma^\star\right).$$

Indeed, if we let $\gamma \to 0$ which corresponds to the case that $\beta_\gamma^\star \to \infty$ and $\theta_\gamma^\star \to \infty$ we have that

$$\lim_{\beta_\gamma^\star, \theta_\gamma^\star \to \infty} \text{GJS}\left(P_1, P_2, \theta_\gamma^\star\right)\text{GJS}\left(P_2, P_1, \beta_\gamma^\star\right) \\ = \text{D}\big(P_2 \| P_1\big)\text{D}\big(P_1 \| P_2\big)$$

This is because of our results in Lemma 4. This shows that our sequential scheme can recover the optimal performance in the case that $\beta_\gamma^\star \to \infty$ and $\theta_\gamma^\star \to \infty$.

C. Comparison to Gutman's Scheme

In this subsection, we compare the proposed sequential test with the Gutman's fixed-length test. We adopt a Bayesian approach in which we assign prior probabilities $\pi_1 \in (0, 1)$ and $\pi_2 = 1 - \pi_1$ to H_1 and H_2 respectively. In this case, the *overall average probability of error* is given by

$$\text{P}^{\text{err}}(\Phi) = \pi_1 \mathbb{P}\left(d\left(X_1^N, X_2^N, Y^T\right) \neq H_1 \middle| H_1\right) \\ + \pi_2 \mathbb{P}\left(d\left(X_1^N, X_2^N, Y^T\right) \neq H_2 \middle| H_2\right) \tag{13}$$

We define the error exponent for the Bayesian scenario as

$$\mathsf{e}_{\text{Bayesian}}^{\pi}\left(\Phi\right) \triangleq \liminf_{N \to \infty} \frac{-\log \text{P}^{\text{err}}(\Phi)}{N}. \tag{14}$$

To make a fair comparison, let us assume each of the schemes, sequential and Gutman, has two training sequences of length N. For Gutman's test, as usual we let $\frac{N}{n} = \alpha$ and we consider the following variation

$$\Phi_{\text{GUT}}(\lambda, \alpha) = \begin{cases} H_1 & \text{if GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \alpha\right) \leq \lambda\alpha, \\ H_2 & \text{if GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \alpha\right) \geq \lambda\alpha, \end{cases} \tag{15}$$

in which λ in (2) is replaced by $\lambda\alpha$ here without loss of generality. For $\Phi_{\text{GUT}}(\lambda, \alpha)$, it can be readily shown that

$$\mathsf{e}^{\pi}_{\text{Bayesian}}\left(\Phi_{\text{GUT}}(\lambda^{\star}, \alpha)\right) = \max_{\lambda \geq 0} \min\left\{\lambda, F_1\left(\alpha, \lambda\right)\right\} \tag{16}$$

where

$$F_1\left(\alpha, \lambda\right) \triangleq \min_{(Q_1, Q_2) \in \mathcal{P}(\mathcal{X})^2} \quad \mathrm{D}\big(Q_1 \| P_1\big) + \frac{1}{\alpha}\mathrm{D}\big(Q_2 \| P_2\big)$$
$$\text{subject to} \quad \frac{1}{\alpha}\text{GJS}\left(Q_1, Q_2, \alpha\right) \leq \lambda. \tag{17}$$

In the next lemma, we study the impact of α on the *Bayesian error exponent* $\mathsf{e}_{\text{Bayesian}}\left(\Phi_{\text{GUT}}\right)$.

Lemma 1: The function $\alpha \mapsto \mathsf{e}^{\pi}_{\text{Bayesian}}\left(\Phi_{\text{GUT}}(\lambda^{\star}, \alpha)\right)$ is decreasing.

Proof: It is straightforward to show that the objective function of (17) is decreasing in α. Also, because $\frac{1}{\alpha}\text{GJS}\left(Q_1, Q_2, \alpha\right)$ is decreasing in α, we conclude that the feasible set is enlarged as α increases. □

Gutman's test is designed for the case that a fixed-length test sequence is provided to the decision maker. On the other hand, from Theorem 2 we know that in the sequential test the average number of the test samples under H_1 and H_2 are different and given approximately by $N/\beta^{\star}_{\gamma}$ and $N/\theta^{\star}_{\gamma}$ respectively. In light of Lemma 1, we will assume, for the sake of comparisons (between Gutman's test and ours), that Gutman's test is provided with $N/\min\{\theta^{\star}_{\gamma}, \beta^{\star}_{\gamma}\}$ test samples, i.e., the (deterministic) number of samples in the testing sequence used by the Gutman's test is equal to the largest expected sample complexity of the sequential test under any of the two hypotheses. We assert that under this assumption, it is fair to compare the sequential and Gutman's tests. The next theorem presents our results concerning the comparison between these two tests.

Theorem 3: Consider the scenario in which $N/\min\{\theta^{\star}_{\gamma}, \beta^{\star}_{\gamma}\}$ samples are available to be used in Gutman's test. Then, achievable Bayesian error exponent of the sequential scheme is *strictly greater* than the Bayesian error exponent of Gutman's test.

Proof: The proof is provided in Appendix IV-C. □

We showed that $\mathsf{e}^{\pi}_{\text{Bayesian}}\left(\Phi_{\text{seq}}(\gamma)\right) \geq \gamma$, and we know $\gamma \in (0, C(P_1, P_2)]$. In the next Corollary we provide the maximum achievable Bayesian error exponent of $\Phi_{\text{seq}}(\gamma)$.

Corollary 2: The maximum achievable Bayesian error exponent of the sequential scheme $\Phi_{\text{seq}}(\gamma)$ is $C\left(P_1, P_2\right)$.

In Figure 2, we provide a numerical example to quantitatively illustrate the gain of our proposed test versus that of the Gutman. We plot an *achievable Bayesian error exponent* based on our sequential scheme compared to the Bayesian error exponent of Gutman's test versus $\min\{\theta^{\star}_{\gamma}, \beta^{\star}_{\gamma}\}$. We consider a ternary alphabet $\mathcal{X} = \{1, 2, 3\}$. In Fig. 2, we set $P_1 = [0.1, 0.3, 0.6]$ and $P_2 = [0.45, 0.45, 0.1]$. As the problem in (17) is a convex problem, we used CVXPY package [21] to perform the optimization. This example shows that sequential test significantly improves the Bayesian error exponent over Gutman's fixed-length test.

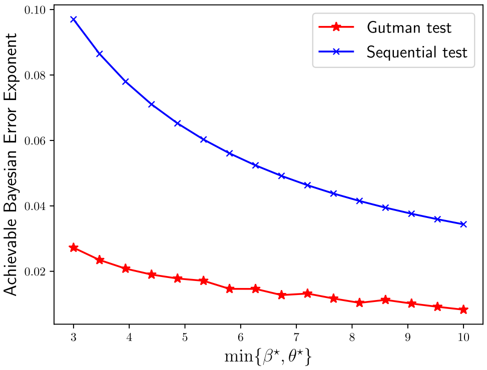

Fig. 2. Comparison of Gutman's test and the sequential test in terms the Bayesian error exponent for $P_1 = [0.1, 0.3, 0.6]$ and $P_2 = [0.45, 0.45, 0.1]$.

III. Sequential Classification: Multi-Class Classification Problem

In this section, we extend the binary classification setup to the scenario in which we have $M > 2$ classes.

A. Problem Statement

In the classification problem with M classes, the decision maker has access to M length-N training sequences denoted by $\{X_i^N\}_{i=1}^{M}$. Each training sequence is generated in an i.i.d. manner according to one of M unknown distributions $(P_1, \ldots, P_M) \in \mathcal{P}\left(\mathcal{X}\right)^M$. We fix a certain distribution P_{i^*} where $i^* \in \{1, \ldots, M\}$. At each time $n \in \mathbb{N}$, a test sample, denoted by Y_n, is generated according to P_{i^*} and is given to the decision maker. The decision maker is tasked to *classify* the test sequence $\{Y_k\}_{k=1}^{n}$, i.e., assign it a label from the set $\{1, \ldots, M\}$. More formally, the decision maker has to decide between the following M hypotheses:

- H_i where $i \in \{1, \ldots, M\}$: The test sequence $\{Y_k\}_{k=1}^{n}$ (which is generated i.i.d. according to P_{i^*}) and the i^{th} training sequence X_i^N are being generated according to the same distribution.

For the described setup, the test, the error probabilities, and the error exponents can be defined analogously to Definitions 1, 2, and 3, respectively. However, here the stopping time T is now adapted to the filtration $\mathcal{F}_n = \sigma\{\{X_i^N\}_{i=1}^{M}, Y_1 \ldots, Y_n\}$, and the terminal decision rule is a function of $\left(\{X_i^N\}_{i=1}^{M}, Y^T\right)$.

The multiclass classification problem has been considered from information-theoretic perspectives in [2]–[6]. There are two main aspects that distinguish our work with previous studies. First, the lengths of the test sequence is fixed in [2]–[6]; however, we let the length of test sequence be random. Second, the M-class classification problem is often studied with the rejection option in the literature; our setup does not include the rejection option. More precisely, in [2]–[6] the decision maker has the following $M + 1$ hypotheses:

- H_i for each $i \in [M]$: The test sequence Y^n and the i^{th} training sequence X_i^N are generated according to the same distribution.
- H_{r}: The test sequence Y^n is generated according to a distribution different from those in which the training sequences are generated from.

Provided that the length of the test sequence n is fixed, in this framework, the error probabilities and the rejection probability are defined as

$$\mathrm{P}_i^{\text{err}}(\Phi) = \mathrm{P}_i\left(d\left(\{X_i^N\}_{i=1}^M, Y^n\right) \notin \{H_i, H_{\text{r}}\}\right) \tag{18}$$

$$\mathrm{P}_i^{\text{rej}}(\Phi) = \mathrm{P}_i\left(d\left(\{X_i^N\}_{i=1}^M, Y^n\right) = H_{\text{r}}\right) \tag{19}$$

for $i \in [M]$. Here note that we are considering *realizable* case in which we know that the test sequence and one of the training sequences are generated according to the same distribution. In (19), $\mathrm{P}_i^{\text{rej}}(\Phi)$ denotes the probability that the decision maker declares H_{r} as the terminal decision (the rejection option is taken here). The main result for the setup with fixed-length test sequence and the rejection option is by Gutman [2]. Note as $N \to \infty$, n also diverges but we have $\lim_{N,n\to\infty} \frac{N}{n} = \alpha$. For example, we can always think of $N = \lfloor n\alpha \rfloor$. In [2] the following questions are addressed: What is the largest $\lambda \in \mathbb{R}$ for which

1) for $i \in [M]$, we have
$$\liminf_{n\to\infty} \frac{-\log \mathrm{P}_i^{\text{err}}(\Phi_{\text{GUT}}^{(M)}(\lambda, \alpha))}{n} \geq \lambda, \tag{20}$$
where $\Phi_{\text{GUT}}^{(M)}(\lambda, \alpha)$ denotes the Gutman's test; and
2) the rejection probability tends to zero as n goes to infinity?

The next theorem provides an answer to the aforementioned question.

Theorem 4: (Gutman [2, Thms. 2 & 3]) Assume $\frac{N}{n} = \alpha$. Then, the maximum λ satisfying both conditions mentioned above is

$$\widehat{\lambda} = \min_{i,j \in [M], i \neq j} \mathrm{GJS}\left(P_i, P_j, \alpha\right). \tag{21}$$

Moreover, if $\lambda > \widehat{\lambda}$, the rejection probability goes to *one* as n goes to infinity.

B. Main Results

In this section, we present our proposed test which does not utilize the rejection option. Let $\gamma \in \mathbb{R}_+$ be a fixed threshold for the test. For $n \in \mathbb{N}$, define the set

$$\Psi_n \triangleq \Bigg\{ i \in \{1, \ldots, M\} : \exists\, 1 \leq k \leq n \text{ such that } k\mathrm{GJS}\left(\widehat{Q}_{X_i^N}, \widehat{Q}_{Y^k}, \frac{N}{k}\right) \geq \gamma N \Bigg\}. \tag{22}$$

Then, the proposed stopping time is

$$T_{\text{seq}}^{(M)} \triangleq \inf\left\{n \geq 1 : |\Psi_n| \geq M - 1\right\} \wedge N^2. \tag{23}$$

Also, at time $T_{\text{seq}}^{(M)}$, the terminal decision rule is

$$d_{\text{seq}}^{(M)} \triangleq [M] \setminus \Psi_{T_{\text{seq}}^{(M)}}. \tag{24}$$

The proposed test is denoted by $\Phi_{\text{seq}}^{(M)}(\gamma) = \left(T_{\text{seq}}^{(M)}, d_{\text{seq}}^{(M)}\right)$.

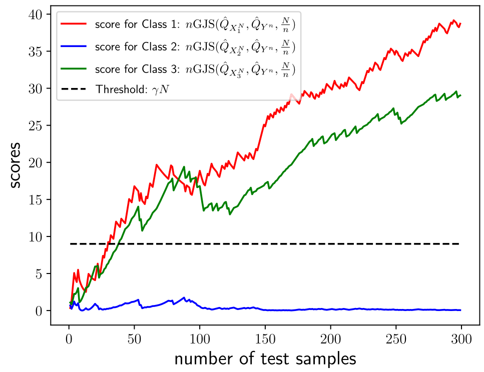

Fig. 3. A realization of the sequential test $\Phi_{\text{seq}}^{(M)}$ for the multi-class classification.

Figure 3 shows a realization of $\Phi_{\text{seq}}^{(M)}(\gamma)$ for three ternary distributions $P_1 = [0.1, 0.7, 0.2]$, $P_2 = [0.4, 0.5, 0.1]$, and $P_3 = [0.3, 0.3, 0.4]$, and $\gamma = 0.03$ where the test sequence drawn from P_2 and the length of each training sequence is $N = 300$. Note that the stopping time defined in (23) is $T_{\text{seq}}^{(M)} = 40$ since at $n = 40$, we have $|\Psi_{73}| = 2 = M - 1$. Then, according to (24), the final decision rule is H_2. We now recall that $C\left(P_i, P_j\right)$, defined in (8), denotes the Chernoff information between P_i and P_j. This will feature in the next theorem.

Theorem 5: Fix $\left(P_1, \ldots, P_M\right) \in \mathcal{P}\left(\mathcal{X}\right)^M$. Let $\mathcal{M} \triangleq \left\{(i,j) \in [M]^2, i \neq j\right\}$. Then, for any $\gamma \in \left[0, \min_{(i,j)\in\mathcal{M}} C\left(P_i, P_j\right)\right]$, the proposed test achieves

- $\mathrm{P}_i^{\text{err}}(\Phi_{\text{seq}}^{(M)}(\gamma)) \stackrel{.}{\leq} \exp\left(-N\gamma\right)$
- $\mathbb{E}_i\left[T_{\text{seq}}^{(M)}\right] = \frac{N}{\min_{j\in[M], j\neq i}\{\theta^\star_{i(j),\gamma}\}}\left(1 + o\left(1\right)\right)$

where $\theta^\star_{i(j),\gamma}$ is given by the solution to the following equation

$$\mathrm{GJS}\left(P_j, P_i, \theta^\star_{i(j),\gamma}\right) = \gamma\theta^\star_{i(j),\gamma}, \qquad \forall\,(i,j) \in \mathcal{M}. \tag{25}$$

Corollary 3: The achievable error exponents of the proposed test under H_i is given by

$$\mathsf{e}_i\left(\Phi_{\text{seq}}^{(M)}(\gamma)\right) = \liminf_{N\to\infty} \frac{-\log \mathrm{P}_i^{\text{err}}(\Phi_{\text{seq}}^{(M)}(\gamma))}{\mathbb{E}_i\left[T_{\text{seq}}^{(M)}\right]} \geq \min_{j\in[M], j\neq i} \mathrm{GJS}\left(P_j, P_i, \theta^\star_{i(j),\gamma}\right), \tag{26}$$

for all $i \in [M]$

C. Comparison to Gutman's Test for the Multiclass Classification Problem

In this section, we compare the Gutman's test for the multiclass classification problem using the test $\Phi_{\text{seq}}^{(M)}$. We adopt a Bayesian approach in which we assign prior probabilities $\pi_i, i \in [M]$ to hypotheses $H_i, i \in [M]$. In this case, the *average probability of error* is given by

$$\mathrm{P}^{\text{err}}(\Phi) = \sum_{i=1}^M \pi_i \mathbb{P}\left(d\left(\{X_i^N\}_{i=1}^M, Y^n\right) \neq H_i \middle| H_i\right). \tag{27}$$

Here, we are interested in the Bayesian error exponent defined in (14). The main result by Gutman in Theorem 4 can be restated in terms of the Bayesian error exponent as follows. The maximum $\lambda_{\text{Bayesian}}$ for which we have

$$\mathsf{e}^{\pi}_{\text{Bayesian}}\left(\Phi^{(M)}_{\text{GUT}}(\lambda,\alpha)\right) \geq \lambda_{\text{Bayesian}},$$

and the rejection probability defined in (19) tends to zero as $N \to \infty$ is

$$\lambda^{\star}_{\text{Bayesian}} = \min_{(i,j)\in\mathcal{M}} \frac{\text{GJS}\left(P_i, P_j, \alpha\right)}{\alpha}. \tag{28}$$

It can be shown that $\lambda^{\star}_{\text{Bayesian}}$ is a decreasing function of α. We follow the same approach as in Section II-C where we compared Gutman's scheme with our proposed test for the binary case. We argue that assigning α to be $\min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j)}$ where $\theta^{\star}_{i(j)}$, defined in (25), ensures that the comparison of the achievable Bayesian error exponents of Gutman's scheme and the sequential test is fair. From Theorem 5, it is straightforward to show that

$$\mathsf{e}^{\pi}_{\text{Bayesian}}\left(\Phi^{(M)}_{\text{seq}}(\gamma)\right) \geq \gamma. \tag{29}$$

Here, we claim that setting $\alpha = \min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j)}$ in (28), we get

$$\begin{aligned}\lambda^{\star}_{\text{Bayesian}} &= \min_{(l,k)\in\mathcal{M}} \frac{\text{GJS}\left(P_l, P_k, \min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j),\gamma}\right)}{\min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j),\gamma}}\\ &= \gamma,\end{aligned} \tag{30}$$

where the last step can be proved as follows: For all $(l,k) \in \mathcal{M}$, we have

$$\text{GJS}\left(P_l, P_k, \min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j),\gamma}\right) \geq \gamma \min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j),\gamma}. \tag{31}$$

The reason for (31) is as follows. Define $f_{l,k}(\theta) \triangleq \text{GJS}\left(P_l, P_k, \theta\right) - \gamma\theta$. Note that $f_{l,k}(\theta^{\star}_{k(l),\gamma}) = 0$. Furthermore, for $\theta \leq \theta^{\star}_{k(l),\gamma}$ we have $f_{l,k}(\theta) \geq \gamma\theta$. Since $\min_{(i,j)\in\mathcal{M}} \theta^{\star}_{i(j),\gamma} \leq \theta^{\star}_{k(l),\gamma}$, we have (31) as required. Therefore, we showed that the Bayesian error exponents of Gutman's test and the sequential test are the same. However, recall that the sequential test achieves this performance *without the rejection option*. The next theorem summarizes our discussion in this section.

Theorem 6: The achievable Bayesian error exponent of the sequential test, i.e., $\Phi^{(M)}_{\text{seq}}(\gamma)$ is equal to that of the Gutman's $\Phi^{(M)}_{\text{GUT}}(\gamma)$. Gutman's test achieves this performance by introducing the rejection option, while the sequential test does not have this option.

Corollary 4: The maximum achievable Bayesian error exponent of the sequential scheme $\Phi^{(M)}_{\text{seq}}(\gamma)$ is

$$\min_{(i,j)\in[M]^2, i\neq j} C\left(P_i, P_j\right), \tag{32}$$

where $C\left(P_i, P_j\right)$ is Chernoff information between P_i and P_j.

Remark 5: It is interesting to note that the possibility of removing the "rejection region" provided that sequential tests are allowed has been shown in other contexts. For instance, in [22, Theorem 6] and [23, Theorem 1] this phenomenon was shown in the context of block coding and streaming data transmission, respectively. However, to the best of our knowledge, none of the previous works consider scenarios in which the distributions are unknown or partially known. This works aims to formalize this idea the more practical statistical learning scenario in which one has partial, noisy information about the underlying distributions in the form of finite-length training samples.

IV. Proofs of the Main Results

A. Preliminary Lemmas

Subsection IV-A is devoted to some preliminary lemmas that will be used in the sequel. We first start with Lemma 2 which provides a variational representation of the GJS divergence.

Lemma 2: Let v^N and w^n are two sequences with types Q_1 and Q_2 over the alphabet $\mathcal{X}$, respectively. We form the following optimization problem.

$$\begin{aligned}\min_{P} \quad & -\frac{1}{n}\log P^{n+N}\left(v^N, w^n\right)\\ \text{subject to} \quad & P \in \mathcal{P}\left(\mathcal{X}\right).\end{aligned} \tag{33}$$

Then, the optimal value of (33) lower bounded by $\text{GJS}\left(Q_1, Q_2, \frac{N}{n}\right)$. Also, the optimal solution is given by $P^{\star} = \frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}$.

Proof: We begin the proof by rewriting

$$N\text{D}\big(Q_1\|P\big) + n\text{D}\big(Q_2\|P\big)$$

$$= N\mathbb{E}_{Q_1}\left[\log\frac{Q_1}{P}\right] + n\mathbb{E}_{Q_2}\left[\log\frac{Q_2}{P}\right] \tag{34}$$

$$= N\mathbb{E}_{Q_1}\left[\log\frac{Q_1}{\frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}}\right] + N\mathbb{E}_{Q_1}\left[\frac{\frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}}{P}\right]$$

$$+ n\mathbb{E}_{Q_2}\left[\log\frac{Q_2}{\frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}}\right] + n\mathbb{E}_{Q_2}\left[\frac{\frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}}{P}\right] \tag{35}$$

$$= n\text{GJS}\left(Q_1, Q_2, \frac{N}{n}\right) + (N+n)\,\text{D}\big(\frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}\|P\big). \tag{36}$$

Then, from [19], we have

$$P^{n+N}\left(v^N, w^n\right) = \exp\left(-N\left(\text{D}\big(Q_1\|P\big)+\text{H}\left(Q_1\right)\right) - n\left(\text{D}\big(Q_2\|P\big)+\text{H}\left(Q_2\right)\right)\right) \tag{37}$$

$$= \exp\left(-n\text{GJS}\left(Q_1, Q_2, \frac{N}{n}\right) - N\text{H}\left(Q_1\right) - n\text{H}\left(Q_2\right) - (N+n)\,\text{D}\big(\frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}\|P\big)\right) \tag{38}$$

$$\leq \exp\left(-n\text{GJS}\left(Q_1, Q_2, \frac{N}{n}\right) - N\text{H}\left(Q_1\right) - n\text{H}\left(Q_2\right)\right) \tag{39}$$

In (38), we plug (34) into (37). Finally, the last step follows due to non-negativity of KL divergence. The upper bound can be achieved by setting $P = \frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}$. Therefore, we can conclude that the optimal solution is $P^{\star} = \frac{\frac{N}{n}Q_1+Q_2}{1+\frac{N}{n}}$. □

Intuitively, if one wishes to find a probability measure that maximizes the joint probability of observing two sequences with different length and *different types*, then GJS naturally arises as a lower bound for the exponent of the desired probability.

In the next lemma, an interesting connection between the GJS divergence and mutual information is established.

Lemma 3: Let X and Y are two independent random variables drawn according to probability distributions P and Q respectively over the same alphabet. Also, W is a Bernoulli random variable independent of X and Y with probabilities $\frac{\alpha}{1+\alpha}$ and $\frac{1}{1+\alpha}$ for $\alpha \in \mathbb{R}_+$, respectively. Let us define the random variable Z as $Z \triangleq \begin{cases} X & \text{if } W = 0 \\ Y & \text{if } W = 1 \end{cases}$. Then, we have

$$(1+\alpha) I\left(Z;W\right) = \mathrm{GJS}\left(P, Q, \alpha\right) \tag{40}$$

Proof: We have

$$\begin{aligned} I\left(Z;W\right) &= \mathrm{H}\left(Z\right) - \mathrm{H}\left(Z\middle|W\right) \\ &= \mathrm{H}\left(Z\right) - \mathrm{H}\left(Z\middle|W=0\right)\frac{\alpha}{1+\alpha} - \mathrm{H}\left(Z\middle|W=1\right)\frac{1}{1+\alpha} \\ &= -\sum_{x\in\mathcal{X}}\left(\frac{\alpha}{1+\alpha}P\left(x\right) + \frac{1}{1+\alpha}Q\left(x\right)\right) \\ &\qquad \times \log\left(\frac{\alpha}{1+\alpha}P\left(x\right) + \frac{1}{1+\alpha}Q\left(x\right)\right) \\ &\quad + \frac{\alpha}{1+\alpha}\sum_{x\in\mathcal{X}} P\left(x\right)\log P\left(x\right) + \frac{1}{1+\alpha}\sum_{x\in\mathcal{X}} Q\left(x\right)\log Q\left(x\right) \\ &= \frac{1}{1+\alpha}\mathrm{GJS}\left(P, Q, \alpha\right) \end{aligned} \tag{41}$$

which gives us the desired result in (40). □

Lemma 4 provides several properties of the GJS divergence.

Lemma 4: For any pair of distributions P and Q in the interior of $\mathcal{P}\left(\mathcal{X}\right)$ and $\alpha \in \mathbb{R}_+$, we have the following facts.

1) $\mathrm{GJS}\left(P, Q, \alpha\right)$ is a concave function in α for fixed P and Q. Moreover, $\lim_{\alpha\to\infty} \mathrm{GJS}\left(P, Q, \alpha\right) = \mathrm{D}\left(Q\|P\right)$.
2) For a fixed $\alpha \in \mathbb{R}_+$, $\mathrm{GJS}\left(P, Q, \alpha\right)$ is a jointly convex function in (P,Q).
3) For fixed P and Q, the necessary and the sufficient condition for the equation $\mathrm{GJS}\left(P, Q, \alpha\right) = \lambda\alpha$ to have a non-zero solution is $\lambda < \mathrm{D}\left(P\|Q\right)$. Also, the solution is unique.

Proof: • Proof of Part (1): To begin with we start by deriving the first and the second derivatives of $\mathrm{GJS}\left(P, Q, \alpha\right)$. We have

$$\frac{\partial \mathrm{GJS}\left(P, Q, \alpha\right)}{\partial \alpha} = \mathrm{D}\big(P\|\frac{\alpha P + Q}{1+\alpha}\big) \tag{42}$$

$$\frac{\partial^2 \mathrm{GJS}\left(P, Q, \alpha\right)}{\partial \alpha^2} = \frac{1}{1+\alpha}\sum_{x\in\mathcal{X}} P\left(x\right)\frac{Q\left(x\right) - P\left(x\right)}{\alpha P\left(x\right) + Q\left(x\right)} \tag{43}$$

We can manipulate the second derivative as

$$\begin{aligned} &\frac{1}{1+\alpha}\sum_{x\in\mathcal{X}} P\left(x\right)\frac{Q\left(x\right) - P\left(x\right)}{\alpha P\left(x\right) + Q\left(x\right)} \\ &= \frac{1}{1+\alpha}\sum_{x\in\mathcal{X}} P\left(x\right)\left(1 - \frac{\left(1+\alpha\right)P\left(x\right)}{\alpha P\left(x\right) + Q\left(x\right)}\right) \\ &= \frac{1}{1+\alpha}\left(1 - \mathbb{E}_P\left[\frac{\left(1+\alpha\right)P\left(X\right)}{\alpha P\left(X\right) + Q\left(X\right)}\right]\right) \\ &\le \frac{1}{1+\alpha}\left(1 - \mathbb{E}_P\left[\frac{\alpha P\left(X\right) + Q\left(X\right)}{\left(1+\alpha\right)P\left(X\right)}\right]^{-1}\right) \qquad (44) \\ &= 0, \end{aligned}$$

where (44) is obtained obtained by applying Jensen's inequality to the convex function $1/x$ for $x > 0$. Thus, we conclude that the second derivative in (43) is negative. Moreover, letting α tend to infinity in (1), we obtain the claim stated in the first part of Lemma 4.

• Proof of Part (2): Consider two pairs of distributions (P_1, Q_1) and (P_2, Q_2). For $0 \le \theta \le 1$, define $P_\theta = \theta P_1 + \left(1-\theta\right)P_2$ and $Q_\theta = \theta Q_1 + \left(1-\theta\right)Q_2$. Then, consider

$$\begin{aligned} &\mathrm{GJS}\left(P_\theta, Q_\theta, \alpha\right) = \\ &\alpha \mathrm{D}\big(\theta \mathrm{P}_1 + \left(1-\theta\right)\mathrm{P}_2 \| \theta\frac{\alpha \mathrm{P}_1 + \mathrm{Q}_1}{1+\alpha} + \left(1-\theta\right)\frac{\alpha \mathrm{P}_2 + \mathrm{Q}_2}{1+\alpha}\big) \\ &+ \mathrm{D}\big(\theta \mathrm{Q}_1 + \left(1-\theta\right)\mathrm{Q}_2 \| \theta\frac{\alpha \mathrm{P}_1 + \mathrm{Q}_1}{1+\alpha} + \left(1-\theta\right)\frac{\alpha \mathrm{P}_2 + \mathrm{Q}_2}{1+\alpha}\big) \\ &\le \theta \mathrm{GJS}\left(P_1, Q_1, \alpha\right) + \left(1-\theta\right)\mathrm{GJS}\left(P_2, Q_2, \alpha\right) \end{aligned}$$

where the last step follows due to the convexity of KL divergence [18, Thm. 2.7.2].

• Proof of Part (3): Let $f\left(\alpha\right) \triangleq \mathrm{GJS}\left(P, Q, \alpha\right) - \lambda\alpha$. First, note that $f\left(0\right) = 0$ and $f\left(\alpha\right)$ is a concave function. It is straightforward to see that $f\left(\alpha\right)$ has at most one non-zero root. Assume that there exists $\alpha^\star > 0$ such that $f\left(\alpha^\star\right) = 0$. By the mean value theorem, we know there exist a $\tilde{\alpha} \in \left(0, \alpha^\star\right)$ such that $f'\left(\tilde{\alpha}\right) = 0$. Knowing this fact and considering the strict concavity of $f\left(\alpha\right)$, we must have $f'\left(0\right) > 0$ which using (42) we have $\lambda < \mathrm{D}\left(P\|Q\right)$. For the other direction, assume that $\lambda < \mathrm{D}\left(P\|Q\right)$. Then, there exist an $\epsilon > 0$ such that for $0 < \alpha < \epsilon$ we have $f\left(\alpha\right) > 0$. Then considering the fact that $\lim_{\alpha\to\infty} f\left(\alpha\right) = -\infty$, f must have a root in the interval $[\epsilon, \infty)$. □

Lemma 5 states an upper bound on the probability that the GJS divergence of two sequences (of different lengths in general) drawn from the same probability distribution exceeds γN.

Lemma 5: Assume X^N and Y^n are two sequences drawn from the *same* probability distribution $P \in \mathcal{P}\left(\mathcal{X}\right)$. Then, we have

$$\mathbb{P}\big(n\mathrm{GJS}\left(\widehat{Q}_{X^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \ge \gamma N\big) \le \exp\left(-\gamma N\right)\left(n+N+1\right)^{|\mathcal{X}|}. \tag{45}$$

Proof: Define $\mathcal{R} = \big\{\left(Q_1, Q_2\right) \in \mathcal{P}_N\left(\mathcal{X}\right) \times \mathcal{P}_n\left(\mathcal{X}\right) \mid n\mathrm{GJS}\left(Q_1, Q_2, \frac{N}{n}\right) \ge \gamma N\big\}$. From the method of types, we can write

$$\begin{aligned} &\mathbb{P}\left(n\mathrm{GJS}\left(\widehat{Q}_{X^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \ge \gamma N\right) \\ &\le \sum_{\left(Q_1, Q_2\right)\in\mathcal{R}} \exp\left(-N\mathrm{D}\left(Q_1\|P\right)\right)\exp\left(-n\mathrm{D}\left(Q_2\|P\right)\right) \qquad (46) \\ &= \sum_{\left(Q_1, Q_2\right)\in\mathcal{R}} \exp\left(-n\mathrm{GJS}\left(Q_1, Q_2, \frac{N}{n}\right)\right) \end{aligned}$$

$$\times \exp\left(- (N+n)\,\mathrm{D}\big(\frac{\frac{N}{n}Q_1 + Q_2}{1+\frac{N}{n}}\|P\big)\right) \tag{47}$$

$$\le \exp(-N\gamma) \sum_{(Q_1,Q_2)\in\mathcal{R}} \exp\big(- (N+n)\,\mathrm{D}\big(\frac{\frac{N}{n}Q_1 + \}Q_2}{1+\frac{N}{n}}\|P\big)\big)$$

$$\le \exp(-N\gamma)\,(N+n+1)^{|\mathcal{X}|} \sum_{(Q_1,Q_2)\in\mathcal{R}} \mathbb{P}\big(\begin{bmatrix} X^N \\ Y^n \end{bmatrix} \in \Gamma^{n+N}_{\frac{NQ_1/n+Q_2}{1+N/n}}\big) \tag{48}$$

$$\le \exp(-N\gamma)\,(N+n+1)^{|\mathcal{X}|}, \tag{49}$$

where the first step is due to the independence of the two sequences and an application of Sanov's theorem. Equation (47) is obtained by using (36) and in (48), we used [18, Theorem 11.1.4] concerning the probability of a type class. □

Lemma 6: Consider two probability distributions $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{X})$. Fix $\gamma > 0$ such that $C(P,Q) > \gamma$ and $C \in \mathbb{R}$ as an arbitrary constant. Denote $\alpha^\star$ as the solution of $\mathrm{GJS}(Q, P, \alpha^\star) = \gamma\alpha^\star$. Let X^N denote a sequence consisting of N i.i.d. samples drawn according to Q. Also, $\alpha^\star_N$ denote the solution of $\mathrm{GJS}\left(\widehat{Q}_{X^N}, P, \alpha^\star_N\right) = \gamma\alpha^\star_N$ if exists, otherwise set $\alpha^\star_N = C$. Then, as N tends to infinity $\alpha^\star_N$ converges in probability to $\alpha^\star$.

Proof: Consider $\epsilon > 0$. Define $\mathcal{S}_N = \{\widehat{Q}_{X^N} \in \mathcal{T}_N \big| \mathrm{D}\big(\widehat{Q}_{X^N}\|P\big) \ge \gamma\}$. From Lemma 4 Part 3, we know that under the event $\mathcal{S}_N$, there exists a solution for $\mathrm{GJS}\left(\widehat{Q}_{X^N}, P, \alpha^\star_N\right) = \gamma\alpha^\star_N$. Then, we can write

$$\begin{aligned} \mathbb{P}(|\alpha^\star_N - \alpha^\star| \ge \epsilon) &= \mathbb{P}(\{|\alpha^\star_N - \alpha^\star| \ge \epsilon\} \cap \{\widehat{Q}_{X^N} \in \mathcal{S}_N\}) \\ &\quad + \mathbb{P}(\{|\alpha^\star_N - \alpha^\star| \ge \epsilon\} \cap \{\widehat{Q}_{X^N} \notin \mathcal{S}_N\}) \\ &\le \mathbb{P}(\{|\alpha^\star_N - \alpha^\star| \ge \epsilon\} \cap \{\widehat{Q}_{X^N} \in \mathcal{S}_N\}) + \mathbb{P}(\widehat{Q}_{X^N} \notin \mathcal{S}_N). \end{aligned} \tag{50}$$

Let $f(V,\alpha) : \mathcal{P}(X) \times \mathbb{R} \to \mathbb{R}$ defined as $f(V,\alpha) \triangleq \mathrm{GJS}(V, P, \alpha) - \gamma\alpha$. First of all note that since $\mathrm{D}\big(P\|Q\big) \ge C(P,Q) > \gamma$, $\alpha^\star$ exists (see Lemma 4 Part 3). Also, note that on the event $\mathcal{S}_N$ we know there exists a solution $\alpha^\star_N$ such that it satisfies $f(\widehat{Q}_{X^N}, \alpha^\star_N) = 0$. From the implicit function theorem [24], since $\frac{\partial f}{\partial \alpha}|_{\alpha=\alpha^\star} \neq 0$ and f is a continuously differentiable function, there exists a unique continuously differentiable function $g : U \to \mathbb{R}$ and a open set U which contains Q such that $g(V) = \alpha$ and $f(V, g(V)) = 0$ for every $V \in U$. Therefore, we can find a sufficiently small $\delta > 0$ such that $\|\widehat{Q}_{X^N} - Q\| \le \delta$ then $|\alpha^\star_N - \alpha^\star| \le \epsilon$. Note that we need to choose δ so that $\{\|\widehat{Q}_{X^N} - Q\| \le \delta\} \subseteq U$. Hence the first term of (50) can be written as

$$\begin{aligned} &\mathbb{P}(\{|\alpha^\star_N - \alpha^\star| \ge \epsilon\} \cap \{\widehat{Q}_{X^N} \in \mathcal{S}_N\}) \\ &\le \mathbb{P}(\{\|\widehat{Q}_{X^N} - Q\| \ge \delta\} \cap \{\widehat{Q}_{X^N} \in \mathcal{S}_N\}). \end{aligned} \tag{51}$$

Due to the fact that $\widehat{Q}_{X^N}$ converges in probability to Q as $N \to \infty$, we conclude that the first term converges to zero as $N \to \infty$. Using the Sanov's theorem, the second term in (50) can be written as

$$\begin{aligned} \mathbb{P}(\widehat{Q}_{X^N} \notin \mathcal{S}_N) &\dot{\le} \exp(-N \min_{V:\mathrm{D}\left(V\|P\right)\le\gamma)} \mathrm{D}\big(V\|Q\big)) \\ &\dot{\le} \exp(-N\gamma) \end{aligned} \tag{52}$$

The reason behind the last line is $\{\gamma \in \mathbb{R} \big| \min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P\right)\le\gamma} \mathrm{D}\big(V\|Q\big) \ge \gamma\} = [0, C(P,Q)]$. Therefore we showed that both term in (50) converges to zero as $N \to \infty$, as was to be shown. □

Lemma 7: Consider the optimization problem

$$\begin{aligned} \min_{(\epsilon_1,\ldots,\epsilon_m)\in\mathbb{R}^m} \quad & \sum_{i=1}^m w_i\epsilon_i \\ \text{subject to} \quad & \sum_{i=1}^m |\epsilon_i| \le \delta, \\ & \sum_{i=1}^m \epsilon_i = 0. \end{aligned} \tag{53}$$

Here, $w_1, \ldots, w_m$ and $\delta > 0$ are constants. Then, the optimal value of the optimization problem in (53) is

$$\frac{\delta}{2}\left(\min_{j\in[m]} w_j - \max_{j\in[m]} w_j\right). \tag{54}$$

Proof: Note that in this proof, the optimal value of variables are denoted using an asterisk in the superscript. We assume without loss of generality that w_i's are all distinct. Otherwise, assume w_{k_1} and w_{k_2} are equal. Assume that we add another constraint to the optimization problem in (53) to get

$$\begin{aligned} \min_{(\epsilon_1,\ldots,\epsilon_m)\in\mathbb{R}^m} \quad & \sum_{i=1}^m w_i\epsilon_i \\ \text{subject to} \quad & \sum_{i=1}^m |\epsilon_i| \le \delta, \\ & \sum_{i=1}^m \epsilon_i = 0 \\ & \epsilon_{k_1} = \epsilon_{k_2}. \end{aligned} \tag{55}$$

We claim that the optimal value of the optimization problem in (55) and (53), denoted by OPT_1 and OPT_2, are equal. The reason is as follows. First of all since the feasible set of (55) is a subset of (53) we conclude that $\mathsf{OPT}_1 \le \mathsf{OPT}_2$. For the other direction, assume $\{\epsilon^\star_{i,1}\}_{i\in[m]}$ are the optimal value of (53). Then consider setting $\epsilon_i = \epsilon^\star_{i,1}$ for $i \notin \{k_1, k_2\}$ and $\epsilon_{k_1} = \epsilon_{k_2} = \frac{1}{2}\epsilon^\star_{k_1,1} + \epsilon^\star_{k_2,1}$. Since $\left|\epsilon^\star_{k_1,1} + \epsilon^\star_{k_1,2}\right| \le \left|\epsilon^\star_{k_1,1}\right| + \left|\epsilon^\star_{k_1,2}\right|$, these values give us a feasible point for (55). Thus, we have $\mathsf{OPT}_1 \ge \mathsf{OPT}_2$. This result shows that given that w_{k_1} and w_{k_2} are equal we can write $w_{k_1}\epsilon_{k_1} + w_{k_2}\epsilon_{k_2} = 2w_{k_1}\epsilon_{k_{1,2}}$ and instead of optimizing over ϵ_{k_1} and ϵ_{k_2} we can only optimize over $\epsilon_{k_{1,2}}$. So, in the sequel, we safely assume all w_i's are distinct. We introduce new variables $\epsilon^+_i \ge 0$ and $\epsilon^-_i \ge 0$ for $i \in [m]$. letting $\epsilon_i = \epsilon^+_i - \epsilon^-_i$, we rewrite (53) as

$$\begin{aligned} \min_{\epsilon_1,\ldots,\epsilon_m} \quad & \sum_{i=1}^m w_i\left(\epsilon^+_i - \epsilon^-_i\right) \\ \text{subject to} \quad & \sum_{i=1}^m \left(\epsilon^+_i + \epsilon^-_i\right) \le \delta, \\ & \sum_{i=1}^m \epsilon^+_i = \sum_{i=1}^m \epsilon^-_i \\ & \epsilon^+_i, \epsilon^-_i \ge 0 \quad \forall i \in [m]. \end{aligned} \tag{56}$$

Note that by $|\epsilon_i| = \epsilon_i^+ + \epsilon_i^-$ and $\epsilon_i = \epsilon_i^+ - \epsilon_i^-$ we implicitly impose the condition $\epsilon_i^+ \epsilon_i^- = 0$ without loss of optimality [25]. The optimization problem in (56) is a linear program, and the optimal solution can be found by considering the Karush-Kuhn-Tucker (KKT) conditions [26, Chapter 5]. Then, we can write the Lagrangian function of (53) as

$$
\begin{aligned}
&\mathcal{L}\left(\left\{\epsilon_i^+\right\}_{i=1}^m, \left\{\epsilon_i^-\right\}_{i=1}^m, \theta, \nu, \left\{\lambda_i^+\right\}_{i=1}^m, \left\{\lambda_i^-\right\}_{i=1}^m\right) \\
&= \sum_{i=1}^m w_i \left(\epsilon_i^+ - \epsilon_i^-\right) + \nu \left(\sum_{i=1}^m \left(\epsilon_i^+ + \epsilon_i^-\right) - \delta\right) \\
&- \sum_{i=1}^m \lambda_i^+ \epsilon_i^+ - \sum_{i=1}^m \lambda_i^- \epsilon_i^- + \theta \sum_{i=1}^m \left(\epsilon_i^+ - \epsilon_i^-\right),
\end{aligned}
$$

where $\nu \geq 0$, $\lambda_i^+ \geq 0$, $\lambda_i^- \geq 0$, and θ are dual variables. Taking the derivatives of Lagrangian with respect to ϵ_i^+ and ϵ_i^- and setting them to zero, we get

$$
\begin{aligned}
w_i + \nu^\star - \left(\lambda_i^+\right)^\star + \theta^\star = 0, \quad \text{and} \qquad (57) \\
-w_i + \nu^\star - \left(\lambda_i^-\right)^\star - \theta^\star = 0, \qquad (58)
\end{aligned}
$$

respectively. Note that given that $w_1, \ldots, w_m$ are not all-zero, it can be verified that there exists at least two indices $i, j \in [m]$ such that we have $\epsilon_i^\star > 0$ and $\epsilon_j^\star < 0$. In this way, for $\epsilon_i^\star$ and $\epsilon_j^\star$, (57) and (58) can be written as

$$
\begin{aligned}
w_i &= -\nu^\star - \theta^\star, \qquad (59) \\
w_j &= \nu^\star - \theta^\star. \qquad (60)
\end{aligned}
$$

In (59) and (60), we have used complementary slackness, i.e., $\lambda_i^{+\star} \epsilon_i^{+\star} = 0$ and $\lambda_j^{-\star} \epsilon_j^{-\star} = 0$. Here we claim that there are exactly two indices for which $|\epsilon_i^\star| > 0$. The reason is that as seen in (59) and (60), $-\nu^\star - \theta^\star$ and $\nu^\star - \theta^\star$ can only take two values. So, given that w_i's are all distinct (59) and (60) can only be satisfied by exactly two indices. Therefore, searching among all $\binom{m}{2}$ combinations of choosing two out of m indices, it is straightforward to see that the optimal solution of (53) is given by

$$
\epsilon_i^\star = \begin{cases} \frac{\delta}{2} & \text{if } i = \arg\min_k w_k, \\ -\frac{\delta}{2} & \text{if } i = \arg\min_k w_k, \\ 0 & \text{else.} \end{cases} \qquad (61)
$$

Thus, the claim stated in the lemma follows. □

B. Proof of the Results for the Binary Case

In this section, we provide the steps toward proving Theorem 2. Specifically, this section consists of three parts. First, we present our results on the expected value of the stopping time. Then, the error probability analysis is provided. Finally, we conclude with the derivation of the error exponent.

For brevity, we present the result for the case that the true hypothesis is H_2, i.e., the underlying probability measure is P_2. The extension of the results here to the case that H_1 is the true hypothesis can be readily done by replacing P_1 by P_2 and vice versa. The following lemma will be used in the the next theorem.

Lemma 8: Assume $f(x) = \mathrm{GJS}\left(\widehat{Q}_{X_1^N}, P_2, x\right) - \gamma x = 0$ has a solution in x and let $\theta_N^\star$ be the solution. Consider

$$
\begin{aligned}
U_+(\theta) \triangleq \max_{V \in \mathcal{P}(\mathcal{X})} \quad & \mathrm{GJS}\left(\widehat{Q}_{X_1^N}, V, \theta\right) \\
\text{subject to} \quad & \mathrm{D}\big(V \| P_2\big) \leq \frac{1}{\sqrt{N}}. \qquad (62)
\end{aligned}
$$

and

$$
\begin{aligned}
U_-(\theta) \triangleq \min_{V \in \mathcal{P}(\mathcal{X})} \quad & \mathrm{GJS}\left(\widehat{Q}_{X_1^N}, V, \theta\right) \\
\text{subject to} \quad & \mathrm{D}\big(V \| P_2\big) \leq \frac{1}{\sqrt{N}}. \qquad (63)
\end{aligned}
$$

Then, for sufficiently large N, we can construct θ_N^+ and θ_N^- which have the following properties.

1) $\theta_N^+ > \theta_N^\star$
2) $\theta_N^+ - \theta_N^\star = O\left(\frac{\log N}{N^{1/4}}\right)$
3) $\theta_N^- < \theta_N^\star$
4) $\theta_N^\star - \theta_N^- = O\left(\frac{1}{N^{1/4}}\right)$

Furthermore, θ_N^+ and θ_N^- satisfy

$$
\begin{aligned}
U_+\left(\theta_N^+\right) < \gamma \theta_N^+, \quad \text{and} \qquad (64) \\
U_-\left(\theta_N^-\right) > \gamma \theta_N^-. \qquad (65)
\end{aligned}
$$

Proof: The proof consists of explicit constructions of θ_N^+ and θ_N^-. The proof is tedious and deferred to Appendix A. □

Before presenting the main properties of the stopping time, we present the following lemma which provides an almost sure lower bound on the stopping time.

Lemma 9: The stopping time defined in (6) is greater than $\left(\frac{\gamma}{2\log 2}\right)^2 N$ almost surely.

Proof: Define

$$
Z_n \triangleq \begin{cases} Z_{n,1} & \text{if } W_n = 1 \\ Z_{n,2} & \text{if } W_n = 2 \end{cases}, \qquad (66)
$$

where $Z_{n,1}$ and $Z_{n,2}$ are distributed according to $\widehat{Q}_{X_1^N}$ and $\widehat{Q}_{Y^n}$, respectively. Also, following the same notation as in Lemma 3, let W_n be a Bernoulli random variable $\mathrm{Bern}(\frac{N}{n+N})$. From the results of Lemma 3, we can write

$$
\begin{aligned}
n\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) &= n\left(1 + \frac{N}{n}\right) I\left(Z_n; W_n\right) \qquad (67) \\
&\leq n\left(1 + \frac{N}{n}\right) \mathrm{H}_b\left(\frac{N}{N+n}\right) \qquad (68) \\
&\leq (n+N)\left(2\log 2\right)\sqrt{\frac{Nn}{(n+N)^2}} \qquad (69) \\
&\leq (2\log 2)\sqrt{nN}.
\end{aligned}
$$

Here, (68) follows due to the fact that $I\left(Z_n; W_n\right) \leq \mathrm{H}_b\left(W_n\right) = \mathrm{H}_b\left(\frac{N}{n+N}\right)$ where $\mathrm{H}_b(p)$ is the binary entropy function defined as $\mathrm{H}_b(p) = -p\log p - (1-p)\log(1-p)$. Finally, in (69) we use $\mathrm{H}_b(p) \leq (2\log 2)\sqrt{p(1-p)}$. Therefore, considering the stopping time in (6), we can conclude that

$$
T_{\text{seq}} \geq \left(\frac{\gamma}{2\log 2}\right)^2 N. \qquad (70)
$$

By replacing $\widehat{Q}_{X_1^N}$ with $\widehat{Q}_{X_2^N}$ in the definition of the random variable in (66), we get the same lower bound as in (70) so the lower bound is agnostic to the true hypothesis. □

Before presenting the next results, we will define a notation. Let X and T are two random variables, and B is a $\sigma(X)$-measurable set. We define $\mathbb{E}[T|X, X \in B]$ as a $\sigma(X)$-measurable function as $\mathbb{E}[T|X, X \in B](x) \triangleq \mathbb{E}[T|X](x)1_B(x)$, where 1_B denotes the indicator function takes value of 1 on B.

Lemma 10: Define the set

$$\mathcal{S}_1 \triangleq \left\{\widehat{Q}_{X_1^N} \in \mathcal{T}_N \,\middle|\, \mathrm{D}\big(\widehat{Q}_{X_1^N}\|P_2\big) \geq \gamma\right\}. \tag{71}$$

The stopping time in (6) has the following properties:

1) $\mathbb{E}_2\big[T_{\text{seq}} \,\big|\, \widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big] \geq \frac{N}{\theta_N^+}\,(1 - o(1))$

2) $\mathbb{E}_2\big[T_{\text{seq}} \,\big|\, \widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big] \leq \frac{N}{\theta_N^-}\,(1 + o(1))$

3) $\mathbb{E}_2\big[T_{\text{seq}}\big] = \frac{N}{\theta^\star}\,(1 + o\,(1))$

where θ_N^+ and θ_N^- are defined in Lemma 8. Also, $\theta^\star$ is the solution of

$$\mathrm{GJS}\,(P_1, P_2, \theta^\star) = \gamma\theta^\star. \tag{72}$$

Proof: First of all, note that for all $\widehat{Q}_{X_1^N} \in \mathcal{S}_1$ (the set $\mathcal{S}_1$ was defined in the statement of Lemma 10), we can assert that there exists a solution to the equation $\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, P_2, \theta_N^\star\right) = \gamma\theta_N^\star$ (See Part 3 of Lemma 4).

Proof of Part 1: We obtain

$$\begin{aligned}
&\mathbb{E}_2\left[T_{\text{seq}}|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right] = \sum_{k\geq 1} \mathrm{P}_2\left(T_{\text{seq}} \geq k \,\middle|\, \widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right)\\
&\geq \sum_{k=1}^{\frac{N}{\theta_N^+}} \mathrm{P}_2\left(T_{\text{seq}} \geq k \,\middle|\, \widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right)\\
&\geq \frac{N}{\theta_N^+}\big(1 - \mathrm{P}_2\big(1 \leq T_{\text{seq}} \leq \frac{N}{\theta_N^+}|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big)\big)\\
&= \frac{N}{\theta_N^+}\big(1 - \mathrm{P}_2\big((\frac{\gamma}{2\log 2})^2\ N \leq T_{\text{seq}} \leq \frac{N}{\theta_N^+}|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big)\big),
\end{aligned} \tag{73}$$

where in (73) we have used Lemma 9. Next, we show that the probability term in (73) is $o(1)$. To do so, we obtain (74) and (75), shown at the bottom of the page. (74) is due to the definition of the stopping time in (6) and the union bound. To obtain the first term in (75), the result of Lemma 5 is used. Then, we provide an upper bound for the second term in (75) as follows. Let $f : \mathbb{N} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be

$$f(k, \widehat{Q}_{X_1^N}) = \min_{V \in \mathcal{P}(\mathcal{X}):k\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, V, \frac{N}{k}\right) \geq \gamma N} \mathrm{D}\big(V\|P_2\big).$$

Then, consider

$$\sum_{k=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta_N^+}} \mathrm{P}_2\big(k\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^k}, \frac{N}{k}\right) \geq N\gamma|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big) \tag{76}$$

$$\leq \sum_{k=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta_N^+}} (k+1)^{|\mathcal{X}|} \exp\big(-kf(k, \widehat{Q}_{X_1^N})\big) \tag{77}$$

$$\leq \sum_{k=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta_N^+}} \big(\frac{N}{\theta_N^+} + 1\big)^{|\mathcal{X}|} \exp\big(-(\frac{\gamma}{2\log 2})^2\ Nf(\frac{N}{\theta_N^+}, \widehat{Q}_{X_1^N})\big) \tag{78}$$

$$\leq \left(\frac{N}{\theta_N^+} + 1\right)^{|\mathcal{X}|+1} \exp\big(-(\frac{\gamma}{2\log 2})^2\sqrt{N}\big), \tag{79}$$

where in (78) we have used the fact that the function $k\mathrm{GJS}\left(\mathrm{P}, \mathrm{Q}, \frac{N}{k}\right)$ is increasing with k. Therefore, as we increase k, the value of the optimization problem in (78) decreases. Finally, the last step comes from the property of θ_N^+ in Lemma 8 which argues that

$$\begin{aligned}
\gamma\theta_N^+ \ \geq \ &\max_{V \in \mathcal{P}(\mathcal{X})} \quad \mathrm{GJS}\left(\widehat{Q}_{X_1^N}, V, \theta_N^+\right)\\
&\text{subject to} \quad \mathrm{D}\big(V\|P_2\big) \leq \frac{1}{\sqrt{N}}.
\end{aligned} \tag{80}$$

This completes the proof of Part 1 of Lemma 10.

Proof of Part 2: We begin with bounding the tail probability of the stopping time. We can write

$$\mathrm{P}_2\left(T_{\text{seq}} > \frac{N}{\theta_N^-}\middle|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right) \tag{81}$$

$$= \sum_{k\geq\frac{N}{\theta_N^-}} \mathrm{P}_2\left(T_{\text{seq}} = k+1\middle|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right) \tag{82}$$

$$\leq \sum_{k\geq\frac{N}{\theta_N^-}} \mathrm{P}_2\left(k\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^k}, \frac{N}{k}\right) \leq \gamma N\middle|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right) \tag{83}$$

$$\begin{aligned}
&\mathrm{P}_2\big((\frac{\gamma}{2\log 2})^2\ N \leq T_{\text{seq}} \leq \frac{N}{\theta_N^+}|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big) \leq \mathrm{P}_2\big(\bigcup_{k=(\frac{\gamma}{2\log 2})^2\ N}^{\frac{N}{\theta_N^+}} \{k\mathrm{GJS}\left(\widehat{Q}_{X_2^N}, \widehat{Q}_{Y^k}, \frac{N}{k}\right) \geq N\gamma\}|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big)\\
&+ \mathrm{P}_2\big(\bigcup_{k=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta_N^+}} \left\{k\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^k}, \frac{N}{k}\right) \geq N\gamma\right\} \middle|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\big)
\end{aligned} \tag{74}$$

$$\leq \frac{N}{\theta_N^+} \exp(-\gamma N)\left(\frac{N}{\theta_N^+} + N + 1\right)^{|\mathcal{X}|} + \sum_{k=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta_N^+}} \mathrm{P}_2\left(k\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^k}, \frac{N}{k}\right) \geq \gamma N\middle|\widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{S}_1\right). \tag{75}$$

$$\leq \sum_{k\geq \frac{N}{\theta_N^-}} (k+1)^{|\mathcal{X}|} \exp\big(-k \min_{V\in\mathcal{P}(\mathcal{X}):k\mathrm{GJS}\left(\widehat{Q}_{X_1^N},V,\frac{N}{k}\right)\leq\gamma N} \mathrm{D}\big(V\|P_2\big)\big) \tag{84}$$

$$\leq \sum_{k\geq \frac{N}{\theta_N^-}} \exp\left(|\mathcal{X}|\log\left(k+1\right)\right)\exp\left(-\frac{k}{\sqrt{N}}\right) \tag{85}$$

$$\leq \exp\left(-\frac{\sqrt{N}}{\theta_N^-}\left(1+o(1)\right)\right). \tag{86}$$

Here, (85) is obtained using the results of Lemma 8 and the fact that $k\mathrm{GJS}\left(\mathrm{P},\mathrm{Q},\frac{N}{k}\right)$ is an increasing function in k. Then the last step follows from some manipulations. Finally, from (82)-(86), we deduce that

$$\mathbb{E}_2\left[T_{\text{seq}}\middle|\widehat{Q}_{X_1^N},\widehat{Q}_{X_1^N}\in\mathcal{S}_1\right]$$

$$\leq \frac{N}{\theta_N^-}\mathrm{P}_2\left(T_{\text{seq}}\leq\frac{N}{\theta_N^-}\middle|\widehat{Q}_{X_1^N},\widehat{Q}_{X_1^N}\in\mathcal{S}_1\right)+$$

$$\sum_{k\geq\frac{N}{\theta_N^-}}(k+1)\,\mathrm{P}_2\left(T_{\text{seq}}=k+1\middle|\widehat{Q}_{X_1^N},\widehat{Q}_{X_1^N}\in\mathcal{S}_1\right) \tag{87}$$

$$\leq \frac{N}{\theta_N^-}\left(1+o(1)\right), \tag{88}$$

which is the desired result.

Proof of Part 3: By the construction of θ_N^+ and θ_N^- in the proof of Lemma 8, when N diverges to infinity, it can be seen that $\theta_N^+-\theta_N^\star=O\left(\frac{\log N}{N^{\frac{1}{4}}}\right)$ and $\theta_N^\star-\theta_N^-=O\left(\frac{1}{N^{\frac{1}{4}}}\right)$. Also, from Lemma 6, we know that $\theta_N^\star$ converges in probability to $\theta_\gamma^\star$. Considering the definition of T_{seq} in (6), we can write

$$\mathbb{E}_2\left[T_{\text{seq}}\right]=\mathbb{E}_2[\mathbb{E}_2[T_{\text{seq}}\,|\,\widehat{Q}_{X_1^N}]\mathbb{1}\{\widehat{Q}_{X_1^N}\in\mathcal{S}_1\}]$$

$$+\,\mathbb{E}_2[\mathbb{E}_2[T_{\text{seq}}\,|\,\widehat{Q}_{X_1^N}]\mathbb{1}\{\widehat{Q}_{X_1^N}\notin\mathcal{S}_1\}] \tag{89}$$

$$=\frac{N}{\theta_\gamma^\star}\left(1+o(1)\right)+N^2\mathrm{P}_2\left(\widehat{Q}_{X_1^N}\in\mathcal{S}_1\right) \tag{90}$$

$$=\frac{N}{\theta_\gamma^\star}\left(1+o(1)\right)+o(1) \tag{91}$$

Here, for the first term of (90) we have used [27, Thm. 2.3.4] to leverage the convergence in probability for $\theta_N^\star\rightarrow\theta^\star$ into convergence in expectation. Note for the final step we have used (114) and Sanov's theorem to write

$$\mathrm{P}_2\left(\widehat{Q}_{X_1^N}\notin\mathcal{S}_1\right)\dot{\leq}\exp\left(-N\gamma\right) \tag{92}$$

and T_{seq} is, almost surely, at most N^2, which is subexponential. □

Corollary 5: When the true hypothesis is H_1, we have

$$\mathbb{E}_1\left[T_{\text{seq}}\right]=\frac{N}{\beta_\gamma^\star}\left(1+o(1)\right), \tag{93}$$

where, $\beta_\gamma^\star$ is the solution of

$$\mathrm{GJS}\left(P_2,P_1,\beta_\gamma^\star\right)=\gamma\beta_\gamma^\star. \tag{94}$$

Therefore, considering the results in Part 3 of Lemma 10 and Corollary 5, the claim in Theorem 2 regarding the stopping time follows immediately.

The following lemma presents bounds on the error probability of the proposed test.

Lemma 11: Under the two different hypotheses, the error probabilities of Φ_{seq} satisfy

$$\mathrm{P}_1^{\text{err}}(\Phi_{\text{seq}}(\gamma))\,\dot{\leq}$$
$$\exp\big(-N\min\big\{\gamma,\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P_1\right)\leq\gamma+\varepsilon_N}\mathrm{D}\big(V\|P_2\big)\big\}\big). \tag{95}$$

$$\mathrm{P}_2^{\text{err}}(\Phi_{\text{seq}}(\gamma))\,\dot{\leq}$$
$$\exp\big(-N\min\big\{\gamma,\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P_2\right)\leq\gamma+\varepsilon'_N}\mathrm{D}\big(V\|P_1\big)\big\}\big). \tag{96}$$

where ε_N and ε'_N are sequences that tend to zero as $N\rightarrow\infty$.

Proof: To compute error probability, we define test $\Phi_{\text{trunc}}(\gamma)$ as a truncated version of $\Phi_{\text{seq}}(\gamma)$. Using $\Phi_{\text{trunc}}(\gamma)$, the decision maker follows the same decision rule as $\Phi_{\text{seq}}(\gamma)$ in the interval $\left[1,N^2\right]$. However, if the stopping time T_{seq} has not occurred in the interval $\left[1,N^2\right]$, the decision maker declares error. It is easy to verify that the error probability of $\Phi_{\text{trunc}}(\gamma)$ is an upper bound for that of $\Phi_{\text{seq}}(\gamma)$. Hence, we can write

$$\mathrm{P}_2^{\text{err}}(\Phi_{\text{seq}}(\gamma))\leq\mathrm{P}_2^{\text{err}}(\Phi_{\text{trunc}}(\gamma)) \tag{97}$$

$$=\mathrm{P}_2\left(\bigcup_{k=1}^{N^2}\left\{n\mathrm{GJS}\left(\widehat{Q}_{X_2^N},\widehat{Q}_{Y^k},\frac{N}{k}\right)\geq\gamma N\right\}\right)$$

$$+\,\mathrm{P}_2\left(T_{\text{seq}}\geq N^2\right) \tag{98}$$

where the first and second term in (98) correspond to the events of "wrong decision" and "no decision" respectively. From Part 3 of Lemma 4, we know that in order to to have a $\theta_N^\star$ which satisfies $\mathrm{GJS}\left(\widehat{Q}_{X_1^N},P_2,\alpha\right)=\gamma\alpha$, we require the condition $\mathrm{D}\big(\widehat{Q}_{X_1^N}\|P_2\big)\geq\gamma$. Also, from the results of Lemma 8 we know that θ_N^- can be constructed using $\theta_N^\star$ given that $\theta_N^\star$ exists. In fact, the map between $\theta_N^\star$ and θ_N^- is one-to-one. Let us define the following set

$$\mathcal{A}_N\triangleq\Bigg\{\widehat{Q}_{X_1^N}\in\mathcal{T}_N\big|\exists\,\theta_N^\star\text{ such that}$$
$$\mathrm{GJS}\left(\widehat{Q}_{X_1^N},P_2,\theta_N^\star\right)=\gamma\theta_N^\star\text{ and }\theta_N^-\geq\frac{N}{N^2}\Bigg\} \tag{99}$$

Next, we argue that since $\theta_N^\star$ is a continuous function of γ, $\mathcal{A}_N$ has another representation which is given by

$$\mathcal{A}_N=\left\{\widehat{Q}_{X_1^N}\in\mathcal{T}_N\,\middle|\,\mathrm{D}\big(\widehat{Q}_{X_1^N}\|P_2\big)\geq\gamma+\varepsilon_N\right\} \tag{100}$$

where $\varepsilon_N\geq 0$ goes to zero as N goes to infinity because as N goes to infinity, $\theta_N^\star$ is greater than zero, and this condition can be satisfied by having $\mathrm{D}\big(\widehat{Q}_{X_1^N}\|P_2\big)>\gamma$ (See Lemma 4).

Then, we can write

$$\begin{aligned}
&\mathrm{P}_2\left(T_{\text{seq}} \geq N^2 \middle| \widehat{Q}_{X_1^N}, \widehat{Q}_{X_1^N} \in \mathcal{A}_N\right) \\
&\leq \sum_{k \geq N^2} \exp\left(|\mathcal{X}| \log(k+1)\right) \exp\left(-\frac{k}{\sqrt{N}}\right) && (101)\\
&\leq \exp\left(-\frac{N^2}{\sqrt{N}}\left(1+o(1)\right)\right) && (102)\\
&\leq \exp\left(-N^{\frac{3}{2}}\left(1+o(1)\right)\right) && (103)
\end{aligned}$$

where in (101) we have used (82)-(85) and the fact that $N^2 \geq \frac{N}{\theta_N^-}$. Then, we obtain

$$\begin{aligned}
&\mathrm{P}_2\left(T_{\text{seq}} \geq N^2\right) \\
&\leq \mathbb{E}_2\left[\mathrm{P}_2\left(T_{\text{seq}} \geq N^2 \middle| \widehat{Q}_{X_1^N}\right) \mathbb{1}\{\widehat{Q}_{X_1^N} \in \mathcal{A}_N\}\right] \\
&\quad + \mathrm{P}_2\left(\mathrm{D}\big(\widehat{Q}_{X_1^N} \| P_2\big) \leq \gamma + \varepsilon_N\right) && (104)\\
&\leq \exp\left(-N^{\frac{3}{2}}\left(1+o(1)\right)\right) + \mathrm{P}_2\left(\mathrm{D}\big(\widehat{Q}_{X_1^N} \| P_2\big) \leq \gamma + \varepsilon_N\right) && (105)\\
&\leq \exp\left(-N^{\frac{3}{2}}\left(1+o(1)\right)\right) \\
&\quad + \exp\left(-N \min_{V \in \mathcal{P}(\mathcal{X}): \mathrm{D}\big(V \| P_2\big) \leq \gamma + \varepsilon_N} \mathrm{D}\big(V \| P_1\big)\right). && (106)
\end{aligned}$$

Here, in (105), we have used (103). Also, the last step follows from Sanov's theorem. Therefore, we obtain

$$\begin{aligned}
&\mathrm{P}_2^{\text{err}}(\Phi_{\text{seq}}(\gamma)) \\
&\leq N^2 \exp\left(-N\gamma\right)\left(N + N^2 + 1\right)^{|\mathcal{X}|} + \exp\left(-N^{\frac{3}{2}}\left(1+o(1)\right)\right) \\
&\qquad + \exp\left(-N \min_{V \in \mathcal{P}(\mathcal{X}): \mathrm{D}\big(V \| P_2\big) \leq \gamma + \varepsilon_N} \mathrm{D}\big(V \| P_1\big)\right) && (107)\\
&\dot{\leq} \exp\left(-N \min\left\{\gamma, \min_{V \in \mathcal{P}(\mathcal{X}): \mathrm{D}\big(V \| P_2\big) \leq \gamma + \varepsilon_N} \mathrm{D}\big(V \| P_1\big)\right\}\right), && (108)
\end{aligned}$$

where the first term in (107) follows by Lemma 5. □

Equipped with the analysis of the stopping time and error probability, we conclude the proof of Theorem 2 by deriving the desired achievable error exponent. We write

$$\begin{aligned}
\mathsf{e}_2\left(\Phi_{\text{seq}}(\gamma)\right) &= \liminf_{N \to \infty} \frac{-\log \mathrm{P}_2^{\text{err}}(\Phi_{\text{seq}}(\gamma))}{\mathbb{E}_2\left[T_{\text{seq}}\right]} && (109)\\
&\geq \theta^\star \liminf_{N \to \infty} \min\left\{\gamma, \min_{V \in \mathcal{P}(\mathcal{X}): \mathrm{D}\big(V \| P_2\big) \leq \gamma + \varepsilon_N} \mathrm{D}\big(V \| P_1\big)\right\} && (110)\\
&= \theta^\star \min\left\{\gamma, \min_{V \in \mathcal{P}(\mathcal{X}): \mathrm{D}\big(V \| P_2\big) \leq \gamma} \mathrm{D}\big(V \| P_1\big)\right\} && (111)\\
&= \theta^\star \gamma && (112)\\
&= \mathrm{GJS}\left(P_1, P_2, \theta_\gamma^\star\right). && (113)
\end{aligned}$$

Here, in (110), we have used Lemmas 10 and 11. The equality in (111) follows from the continuity of the optimal value of the optimization problem with respect to ε_N [26, Sec 5.6]. In (112), we used the fact that

$$\left\{\gamma \,\middle|\, \min_{V \in \mathcal{P}(\mathcal{X}): \mathrm{D}\big(V \| P_2\big) \leq \gamma} \mathrm{D}\big(V \| P_1\big) \geq \gamma\right\} = [0, C\left(P_1, P_2\right)]. \tag{114}$$

Finally, the last step in (113) is obtained due to the defintion of $\theta_\gamma^\star$ in (10). Note that the extension of the results here to the type-I error exponent can be readily done which leads to the statement in Theorem 2.

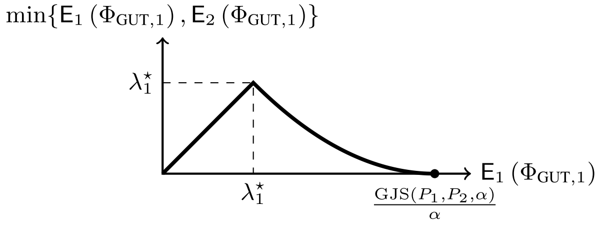

Fig. 4. The performance of the Gutman's test when the *first* training sequence is used.

C. Proof of Theorem 3

In this part, we denote $\Phi_{\text{GUT},1}$ to denote the test described in (15). The subscript 1 in $\Phi_{\text{GUT},1}$ represents the fact that the test uses the first training sequence. In this subsection, we prove Theorem 3 which states that the proposed test outperforms the Gutman's test in terms of Bayesian error exponent defined in (14). We first begin with proving a property of the Gutman's test which will be used in the main proof.

For the test described in (15), we have

$$\begin{aligned}
\mathsf{E}_1\left(\Phi_{\text{GUT},1}\right) &\triangleq \liminf_{N \to \infty} \frac{-\log \mathrm{P}_1^{\text{err}}(\Phi_{\text{GUT},1})}{N} \geq \lambda, \quad \text{and} && (115)\\
\mathsf{E}_2\left(\Phi_{\text{GUT},1}\right) &\triangleq \liminf_{N \to \infty} \frac{-\log \mathrm{P}_2^{\text{err}}(\Phi_{\text{GUT},1})}{N} \geq F_1\left(\alpha, \lambda\right). && (116)
\end{aligned}$$

A schematic of $\min\{\mathsf{E}_1\left(\Phi_{\text{GUT},1}\right), \mathsf{E}_2\left(\Phi_{\text{GUT},1}\right)\}$ versus λ is depicted in Figure 4. Two important observations are in order.

- For $\lambda \geq \frac{1}{\alpha}\mathrm{GJS}\left(P_1, P_2, \alpha\right)$, we have $\min\{\mathsf{E}_1\left(\Phi_{\text{GUT},1}\right), \mathsf{E}_2\left(\Phi_{\text{GUT},1}\right)\} = 0$ as a consequence of (17) .
- $\lambda_1^\star$ in Fig. 4 denotes the maximum achievable Bayesian error exponent as defined in (16).

It is important to note that although two training sequences are produced, only one of them X_1^N is used in (15). One can suggest the following test which resembles the one in (15) but uses the second training sequence as

$$\Phi_{\text{GUT,2}} = \begin{cases} H_2 & \text{if } \mathrm{GJS}\left(\widehat{Q}_{X_2^N}, \widehat{Q}_{Y^n}, \alpha\right) \leq \lambda\alpha, \\ H_1 & \text{if } \mathrm{GJS}\left(\widehat{Q}_{X_2^N}, \widehat{Q}_{Y^n}, \alpha\right) \geq \lambda\alpha, \end{cases}. \tag{117}$$

Note that the $\Phi_{\text{GUT,1}}$ and $\Phi_{\text{GUT,2}}$ depend on α and λ, but we do not want to show the dependence due to the notational convenience. The extension of the Gutman's main theorem to the test in (117) is given by the following lemma.

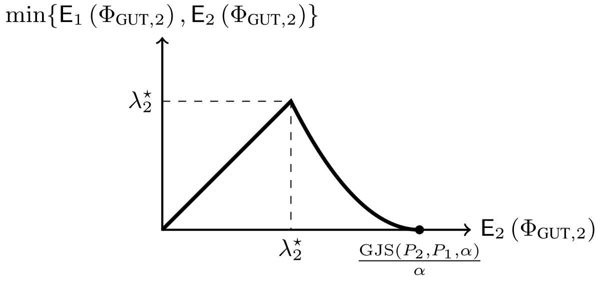

Fig. 5. The performance of the Gutman's test when the *second* training sequence is used.

Lemma 12: Among all decision rules Φ such that for all pairs of distribution $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$,

$$\liminf_{N\to\infty} \frac{-\log \mathrm{P}_2^{\text{err}}(\Phi_{\text{GUT},2}(\lambda,\alpha))}{N} \geq \lambda, \tag{118}$$

the test $\Phi_{\text{GUT},2}$ in (117) satisfies

$$\liminf_{N\to\infty} \frac{-\log \mathrm{P}_1^{\text{err}}(\Phi_{\text{GUT},2})}{N} \geq \liminf_{N\to\infty} \frac{-\log \mathrm{P}_1^{\text{err}}(\Phi)}{N}. \tag{119}$$

Also, given that $\alpha = \frac{N}{n}$, we obtain

$$\mathsf{E}_1(\Phi_{\text{GUT},2}) = \liminf_{N\to\infty} \frac{-\log \mathrm{P}_1^{\text{err}}(\Phi_{\text{GUT},2})}{N} \geq F_2(\alpha,\lambda), \tag{120}$$

$$\mathsf{E}_2(\Phi_{\text{GUT},2}) = \liminf_{N\to\infty} \frac{-\log \mathrm{P}_2^{\text{err}}(\Phi_{\text{GUT},2})}{N} \geq \lambda, \tag{121}$$

where

$$\begin{aligned} F_2(\alpha,\lambda) \triangleq & \min_{(V_1,V_2)\in\mathcal{P}(\mathcal{X})^2} \quad \mathrm{D}(V_1\|P_2) + \frac{1}{\alpha}\mathrm{D}(V_2\|P_1) \\ & \text{subject to} \quad \frac{1}{\alpha}\text{GJS}(V_1,V_2,\alpha) \leq \lambda. \end{aligned} \tag{122}$$

In Figure 5, we show a schematic plot of $\min\{\mathsf{E}_1(\Phi_{\text{GUT},2}), \mathsf{E}_2(\Phi_{\text{GUT},2})\}$ versus $\mathsf{E}_2(\Phi_{\text{GUT},2})$. Similar to Figure 4, we observe that

- For $\lambda \geq \frac{1}{\alpha}\text{GJS}(P_2,P_1,\alpha)$, we have $\min\{\mathsf{E}_1(\Phi_{\text{GUT},2}), \mathsf{E}_2(\Phi_{\text{GUT},2})\} = 0$.
- Also, $\lambda_2^\star$ in Fig. 5 depicts the maximum achievable Bayesian error exponent of $\Phi_{\text{GUT},2}$.

Lemma 13: The maximum achievable Bayesian error exponents of $\Phi_{\text{GUT},1}$ and $\Phi_{\text{GUT},2}$ are equal.
Hence, the Bayesian error exponent of Gutman's test is agnostic to which training sequence is being used.

Proof: Here, we want to prove that $\lambda_1^\star = \lambda_2^\star$. The proof is by contradiction. Assume that $\lambda_1^\star < \lambda_2^\star$. Consider the tradeoff of type-I and type-II error exponents in Figure 5. Then, denote λ^+ as the solution to $F_2(\alpha,\lambda^+) = \lambda_1^\star$. Since $\lambda_1^\star < \lambda_2^\star$ and $F_2(\alpha,\lambda)$ is decreasing function in λ, it can be verified that $\lambda^+ \in \left(\lambda_2^\star, \frac{1}{\alpha}\text{GJS}(P_2,P_1,\alpha)\right)$. Therefore, we have $\lambda^+ > \lambda_2^\star > \lambda_1^\star$. Here, we want to prove that λ^+ being greater than $\lambda_1^\star$ contradicts with optimality of Gutman's test described in Theorem 1. Assume that $\lambda^+ > \lambda_1^*$, we can argue that the test based on the second training sequence achieves the type-I error exponent equal to $\lambda_1^\star$ while its type-II error exponent is $\lambda^+ > \lambda_1^\star$. This contradicts with the fact that among all tests that achieve the same type-I error exponent, Gutman's test has the largest type-II exponent. By the same argument, it can be shown that $\lambda_2^* < \lambda_1^*$ contradicting Lemma 12. Thus, we have $\lambda_1^\star$ in Fig. 4 is equal to $\lambda_2^\star$ in Fig. 5. □

Now, the result of Lemma 13 allows us to prove Theorem 3. Consider the following two scenarios separately.

1) $\theta_\gamma^\star \leq \beta_\gamma^\star$: Given that $\theta^\star \leq \beta^\star$, we have $\gamma = \frac{1}{\theta^\star}\text{GJS}(P_1,P_2,\theta^\star)$ as shown in Theorem 2, and γ is the maximum achiavable exponent of $\Phi_{\text{seq}}(\gamma)$. Considering $\alpha = \theta^\star$ for Gutman's test,

$$\lambda_1^\star \overset{(a)}{<} \text{GJS}(P_1,P_2,\alpha)/\alpha \overset{(b)}{=} \text{GJS}\left(P_1,P_2,\theta_\gamma^\star\right)/\theta_\gamma^\star \overset{(c)}{=} \gamma.$$

Here, (a) is by Figure 4, (b) follows since $\alpha = \min\{\theta_\gamma^\star, \beta_\gamma^\star\} = \theta_\gamma^\star$, and (c) is due to Theorem 2.

2) $\theta_\gamma^\star > \beta_\gamma^\star$: In this case, for the sequential test we have $\gamma = \frac{1}{\beta^\star}\text{GJS}(P_2,P_1,\beta^\star)$, and .

$$\lambda_2^\star \overset{(a)}{<} \text{GJS}(P_2,P_1,\alpha)/\alpha \overset{(b)}{=} \text{GJS}\left(P_2,P_1,\beta_\gamma^\star\right)/\beta_\gamma^\star \overset{(c)}{=} \gamma.$$

Here, (a) is by Figure 5, (b) follows since $\alpha = \min\{\theta_\gamma^\star, \beta_\gamma^\star\} = \beta_\gamma^\star$, and (c) is due to Theorem 2. Then, using the fact that $\lambda_2^\star = \lambda_1^\star = \mathsf{e}_{\text{Bayesian}}^\pi(\Phi_{\text{GUT}}(\lambda^\star,\alpha))$, the claim stated in Theorem 3 is proved.

D. Proof of the Results for Multi-Class Classification Problem

This section consists of three parts: stopping time analysis, derivation of the error probability, and finally characterizing the achievable error exponent.

Our main result on the expected value $T_{\text{seq}}^{(M)}$ is presented in the next lemma.

Lemma 14: Denote $\theta_{i(j),\gamma}^\star$ as the solution of the equation

$$\text{GJS}\left(P_j,P_i,\theta_{i(j),\gamma}^\star\right) = \gamma\theta_{i(j),\gamma}^\star, \quad j \in [M], i \neq j. \tag{123}$$

Then, the expected value of $T_{\text{seq}}^{(M)}$ satisfies

$$\mathbb{E}_i\left[T_{\text{seq}}^{(M)}\right] = \frac{N}{\min_{j\in[M],j\neq i}\{\theta_{i(j),\gamma}^\star\}}(1+o(1)), \tag{124}$$

for all $i \in \{1,\ldots,M\}$.

Proof: Let us assume that the test sequence generated from P_1, i.e., belongs to Class 1. The extension to other cases is straightforward. Define the set

$$\mathcal{S}_1^{(M)} \triangleq \left\{\left(\widehat{Q}_{X_2^N},\ldots,\widehat{Q}_{X_M^N}\right) \middle| \left(\widehat{Q}_{X_2^N},\ldots,\widehat{Q}_{X_M^N}\right) \in \prod_{i=2}^{M}\left\{\widehat{Q}_{X_i^N} \in \mathcal{T}_N \,\middle|\, \mathrm{D}\left(\widehat{Q}_{X_i^N}\|P_1\right) \geq \gamma\right\}\right\} \tag{125}$$

Conditioned on $\mathcal{S}_1^{(M)}$, we can find $\theta_{N,1(j)}^\star$ such that $\theta_{N,1(j)}^\star$ satisfies

$$\text{GJS}\left(\widehat{Q}_{X_j^N},P_1,\theta_{N,1(j)}^\star\right) = \gamma\theta_{N,1(j)}^\star \quad j \in \{2,\ldots,M\}. \tag{126}$$

Also, define

$$\theta^{\star}_{N,1} \triangleq \min_{j\in\{2,\dots,M\}} \{\theta^{\star}_{N,1(j)}\}, \quad \text{and} \tag{127}$$

$$j^{\star} \triangleq \operatorname*{arg\,min}_{j\in\{2,\dots,M\}} \{\theta^{\star}_{N,1(j)}\}. \tag{128}$$

In addition, we substitute P_2 with P_1 and $\widehat{Q}_{X_1^N}$ with $\widehat{Q}_{X_{j^\star}^N}$ in Lemma 8 to obtain $\theta^{+}_{N,1(j^\star)}$ and $\theta^{-}_{N,1(j^\star)}$ following the same procedure as described in Lemma 8. We start with providing a lower bound on the expected value of the stopping time. We can write

$$\begin{aligned}
&\mathbb{E}_1\left[T_{\text{seq}}^{(M)}\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\right] \\
&= \sum_{k=1}^{\infty} \mathrm{P}_1\left(T_{\text{seq}}^{(M)} \geq k\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\right) && (129)\\
&\geq \sum_{k=1}^{\frac{N}{\theta^{+}_{N,1(j^\star)}}} \mathrm{P}_1\left(T_{\text{seq}}^{(M)} \geq k\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\right) && (130)\\
&\geq \frac{N}{\theta^{+}_{N,1(j^\star)}}\big(1-\mathrm{P}_1\big(1\leq T_{\text{seq}}^{(M)} \leq \frac{N}{\theta^{+}_{N,1(j^\star)}}\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\big)\big) && (131)\\
&= \frac{N}{\theta^{+}_{N,1(j^\star)}}\big(1- \\
&\mathrm{P}_1\big((\frac{\gamma}{2\log 2})^2\ N \leq T_{\text{seq}}^{(M)} \leq \frac{N}{\theta^{+}_{N,1(j^\star)}}\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\big)\big) && (132)
\end{aligned}$$

where in the last step we have used Lemma 9. Moreover, define

$$\tau_i^{(M)} = \inf\left\{n\geq 1\ :\ n\mathrm{GJS}\left(\widehat{Q}_{X_i^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \geq \gamma N\right\} \tag{133}$$

as the time that empirical GJS divergence between the test sequences and the i-th training sequence exceeds the threshold. Then, we upper bound the probability term in (132) shown at the bottom of the page in (134)-(137).

Here, the first term on the LHS of (136) is obtained by the same reasons as those for (76)-(79). Also, the second term on the LHS of (136) follows from Lemma 5. Thus, we conclude from (132) and (137) that

$$\mathbb{E}_1\left[T_{\text{seq}}^{(M)}\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\right] \geq \frac{N}{\theta^{+}_{N,1(j^\star)}}\left(1-o\left(1\right)\right). \tag{138}$$

At the bottom of the next page in (139)-(142), an upper bound on the tail probability of $T_{\text{seq}}^{(M)}$ is derived which leads to an upper bound on the expected value of $T_{\text{seq}}^{(M)}$. Here, (140) is obtained using the fact that the event $\{T_{\text{seq}}^{(M)} = k+1\}$ has the same probability as the event that there exists at least two indices $(i,j) \in [M]^2$ such that $\tau_i^{(M)} > k$ and $\tau_j^{(M)} > k$. Then, (142) follows from the definition of $j^\star$ in (128) which attains the minima. Then, following the same line of reasoning as (88) we obtain

$$\mathbb{E}_1\left[T_{\text{seq}}^{(M)}\ \Big|\ \{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\right] \leq \frac{N}{\theta^{-}_{N,1(j^\star)}}\left(1+o\left(1\right)\right) \tag{143}$$

Finally, note that as it was proved in Lemma 6, $\theta^{\star}_{N,1(j)}$ converges in probability to $\theta^{\star}_{1(j)}$ for $j \in \{2,\dots,M\}$ as N goes to infinity. Also, since min function is continuous in its argument, the continuous mapping theorem [28] implies that $\theta^{\star}_{N,1}$ converges in probability to $\min_{j\in[M],j\neq 1}\{\theta^{\star}_{1(j),\gamma}\}$.

$$\begin{aligned}
&\mathrm{P}_1\big((\frac{\gamma}{2\log 2})^2\ N \leq T_{\text{seq}}^{(M)} \leq \frac{N}{\theta^{+}_{N,1(j^\star)}}\Big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\big) \\
&\leq \mathrm{P}_1\big((\frac{\gamma}{2\log 2})^2\ N \leq \max\{\tau_2^{(M)},\dots,\tau_M^{(M)}\} \leq \frac{N}{\theta^{+}_{N,1(j^\star)}}, \bigcap_{n=\left(\frac{\lambda}{2\log 2}\right)^2\ N}^{\frac{N}{\theta^{+}_{N,1(j^\star)}}} \left\{n\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \leq \gamma N\right)\big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\big) \\
&\quad + \mathrm{P}_1\big(\bigcup_{n=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta^{+}_{N,1(j^\star)}}} \left\{n\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \geq \gamma N\right)\ \big|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\big) && (134)\\
&\leq \mathrm{P}_1\left(\left(\frac{\gamma}{2\log 2}\right)^2\ N \leq \tau_{j^\star}^{(M)} \leq \frac{N}{\theta^{+}_{N,1(j^\star)}}\Bigg|\{\widehat{Q}_{X_j^N}\}_{2\leq j\leq M}, \mathcal{S}_1^{(M)}\right) + \mathrm{P}_1\big(\bigcup_{n=\left(\frac{\gamma}{2\log 2}\right)^2\ N}^{\frac{N}{\theta^{+}_{N,1(j^\star)}}} \left\{n\mathrm{GJS}\left(\widehat{Q}_{X_1^N}, \widehat{Q}_{Y^n}, \frac{N}{n}\right) \geq \gamma N\right\}\big) && (135)\\
&\leq \left(\frac{N}{\theta^{+}_{N,1(j^\star)}}+1\right)^{|\mathcal{X}|+1} \exp\left(-\left(\frac{\gamma}{2\log 2}\right)^2\sqrt{N}\right) + \frac{N}{\theta^{+}_{N,1(j^\star)}}\exp\left(-\gamma N\right)\left(\frac{N}{\theta^{+}_{N,1(j^\star)}}+N+1\right)^{|\mathcal{X}|} && (136)\\
&= o\left(1\right). && (137)
\end{aligned}$$

To conclude the proof, we write

$$\begin{aligned}&\mathbb{E}_1\left[T_{\text{seq}}^{(M)}\right] = \\ &\mathbb{E}_1[\mathbb{E}_1[T_{\text{seq}}^{(M)} \,|\, \{\widehat{Q}_{X_j^N}\}_{2\le j\le M}]\mathbb{1}\{\{\widehat{Q}_{X_j^N}\}_{2\le j\le M} \in \mathcal{S}_1^{(M)}\}]+ \\ &\mathbb{E}_1[\mathbb{E}_1[T_{\text{seq}} \,|\, \{\widehat{Q}_{X_j^N}\}_{2\le j\le M}]\mathbb{1}\{\{\widehat{Q}_{X_j^N}\}_{2\le j\le M} \notin \mathcal{S}_1^{(M)}\}] \qquad (144) \\ &\le \frac{N}{\min_{j\in[M],j\neq 1}\{\theta^\star_{1(j),\gamma}\}}(1+o(1))+N^2\mathrm{P}_1\left(\widehat{Q}_{X_1^N}\in\mathcal{S}_1^{(M)}\right) \qquad (145) \\ &\le \frac{N}{\min_{j\in[M],j\neq 1}\{\theta^\star_{1(j),\gamma}\}}(1+o(1))+o(1) \qquad (146)\end{aligned}$$

Note in the second term of the final step we have used (161) to write

$$1-\mathrm{P}_1\left(\{\widehat{Q}_{X_j^N}\}_{2\le j\le M}\in\mathcal{S}_1^{(M)}\right) \dot{\le} \exp\left(-N\gamma\right) \qquad (147)$$

and T_{seq}^M is, almost surely, at most N^2, which is subexponential. □

Lemma 15: The error probability of the test $\Phi_{\text{seq}}^{(M)}(\gamma)$ is given by

$$\begin{aligned}&\mathrm{P}_i^{\text{err}}(\Phi_{\text{seq}}^{(M)}(\gamma)) \dot{\le} \\ &\exp\big(-N\min\big\{\gamma, \min_{j\in[M],j\neq i}\ \min_{V:\mathrm{D}\left(V\|P_j\right)\le\gamma+\varepsilon_{j,N}} \mathrm{D}\big(V\|P_i\big)\big\}\big), \qquad (148)\end{aligned}$$

where $\varepsilon_{j,N} \ge 0$ is a sequence for each $j \in \{1,...,M\}$ converging to zero as N tends to infinity for all $i \in [M]$.

Proof: We define a test $\Phi_{\text{trunc}}^{(M)}$ to be a truncated version of $\Phi_{\text{seq}}^{(M)}$ in an exactly similar way as we defined $\Phi_{\text{trunc}}^{(M)}$ in the proof of Lemma 11. Then, we can write

$$\begin{aligned}\mathrm{P}_1^{\text{err}}(\Phi_{\text{seq}}^{(M)}) &\le \mathrm{P}_1^{\text{err}}(\Phi_{\text{trunc}}^{(M)}) \qquad (149) \\ &\le \mathrm{P}_1\left(\bigcup_{n=1}^{N^2}\left\{n\mathrm{GJS}\left(\widehat{Q}_{X_1^N},\widehat{Q}_{Y^n},\frac{N}{n}\right)\ge\gamma N\right\}\right) \\ &\quad + \mathrm{P}_1\left(T_{\text{seq}}^{(M)}\ge N^2\right). \qquad (150)\end{aligned}$$

Note that the first and the second term in (150) correspond to the event “wrong decision" and the “no decision". Following the same line of reasoning as in the proof of Lemma 11, we consider the event $\bigcap_{i=2}^{M}\left\{\mathrm{D}\big(\widehat{Q}_{X_i^N}\|P_1\big) > \gamma+\varepsilon_{i,N}\right\}$ where $\varepsilon_{i,N} \ge 0$ is a sequence goes to zero as N goes to infinity. Conditioned on this event we can conclude that there exists $\theta^\star_{N,1(i)}$ which satisfies equation $\mathrm{GJS}\left(\widehat{Q}_{X_i^N}, P_1, \theta^\star_{N,1(i)}\right) = \gamma\theta^\star_{N,1(i)}$ for $i \in \{2,\ldots,M\}$ by Part 3 of Lemma 4 . Define $\theta^-_{N,1(i^\star)}$ following the method described in Lemma 8. Introducing $\varepsilon_{i,N}$ let us have $N^2 \ge N/\theta^-_{N,1(i)}$ for $i \in \{2,\ldots,M\}$. Also, let $\theta^\star_{N,1(i^\star)} \triangleq \min_{i\in\{2,\ldots,M\}}\{\theta^\star_{N,1(i)}\}$. Then, we can write

$$\begin{aligned}&\mathrm{P}_1\big(T_{\text{seq}}^{(M)}\ge N^2\big|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M},\bigcap_{i=2}^{M}\big\{\mathrm{D}\big(\widehat{Q}_{X_i^N}\|P_1\big)>\gamma+\varepsilon_{i,N}\big\}\big) \\ &\le \sum_{k\ge N^2}\exp\left(|\mathcal{X}|\log\left(k+1\right)\right)\exp\left(-\frac{k}{\sqrt{N}}\right) \qquad (151) \\ &\le \frac{M\left(M-1\right)}{2}\exp\left(-\frac{N^2}{\sqrt{N}}\left(1+o\left(1\right)\right)\right) \qquad (152) \\ &= \frac{M\left(M-1\right)}{2}\exp\left(-N^{\frac{3}{2}}\left(1+o\left(1\right)\right)\right) \qquad (153)\end{aligned}$$

where in (151) we have used (139)-(142) and the fact that $\theta^-_{N,1(i^\star)} \ge N/N^2$. We obtain

$$\begin{aligned}&\mathrm{P}_1\left(T_{\text{seq}}^{(M)}\ge N^2\right) \le \\ &\mathbb{E}_1\big[\mathrm{P}_1\left(T_{\text{seq}}^{(M)}\ge N^2\middle|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M}\right) \times \\ &\qquad \mathbb{1}\{\bigcap_{i=2}^{M}\left\{\mathrm{D}\big(\widehat{Q}_{X_i^N}\|P_1\big)>\gamma+\varepsilon_{i,N}\right\}\}\big] \\ &\quad + \mathrm{P}_1\left(\bigcup_{i=2}^{M}\left\{\mathrm{D}\big(\widehat{Q}_{X_i^N}\|P_1\big)\le\gamma+\varepsilon_{i,N}\right\}\right) \\ &\le \frac{M\left(M-1\right)}{2}\exp\left(-N^{\frac{3}{2}}\left(1+o\left(1\right)\right)\right)+\end{aligned}$$

$$\begin{aligned}&\mathrm{P}_1\left(T_{\text{seq}}^{(M)} > \frac{N}{\theta^-_{N,1(j^\star)}}\middle|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M},\mathcal{S}_1^{(M)}\right) = \sum_{k=\frac{N}{\theta^-_{N,1(j^\star)}}}^{\infty}\mathrm{P}_1\left(T_{\text{seq}}^{(M)}=k+1\middle|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M},\mathcal{S}_1^{(M)}\right) \qquad (139) \\ &\le \sum_{k=\frac{N}{\theta^-_{N,1(j^\star)}}}^{\infty}\ \sum_{(i_1,i_2)\in[M]^2,i_1\neq i_2}\mathrm{P}_1\left(\tau_{i_1}^{(M)}>k,\tau_{i_2}^{(M)}>k\middle|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M},\mathcal{S}_1^{(M)}\right) \qquad (140) \\ &\le \sum_{(i_1,i_2)\in[M]^2,i_1\neq i_2,i_1=1}\ \sum_{k=\frac{N}{\theta^-_{N,1(j^\star)}}}^{\infty}\mathrm{P}_1\left(k\mathrm{GJS}\left(\widehat{Q}_{X_{i_2}^N},\widehat{Q}_{Y^k},\frac{N}{k}\right)\le\gamma N\middle|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M},\mathcal{S}_1^{(M)}\right) \\ &\quad + \sum_{(i_1,i_2)\in[M]^2,i_1\neq i_2,i_1\neq 1}\ \sum_{k=\frac{N}{\theta^-_{N,1(j^\star)}}}^{\infty}\mathrm{P}_1\left(k\mathrm{GJS}\left(\widehat{Q}_{X_{i_1}^N},\widehat{Q}_{Y^k},\frac{N}{k}\right)\le\gamma N\middle|\{\widehat{Q}_{X_j^N}\}_{2\le j\le M},\mathcal{S}_1^{(M)}\right) \qquad (141) \\ &\le \frac{M\left(M-1\right)}{2}\exp\left(-\frac{\sqrt{N}}{\theta^-_{N,1(j^\star)}}\left(1+o(1)\right)\right) \qquad (142)\end{aligned}$$

$$\sum_{i=2}^{M}\exp\left(-N\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P_1\right)\le\gamma+\varepsilon_{i,N}}\mathrm{D}\big(V\|P_i\big)\right) \tag{154}$$

Plugging (154) into (150), we get

$$\begin{aligned}
&\mathrm{P}_1^{\mathrm{err}}(\Phi_{\mathrm{seq}}^{(M)}(\gamma)) \le N^2\exp\left(-\gamma N\right)\left(N+N^2+1\right)^{|\mathcal{X}|}\\
&+\frac{M\left(M-1\right)}{2}\exp\left(-N^{\frac{3}{2}}\left(1+o\left(1\right)\right)\right)\\
&\quad+\sum_{i=2}^{M}\exp\left(-N\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P_1\right)\le\gamma+\varepsilon_{i,N}}\mathrm{D}\big(V\|P_i\big)\right)
\end{aligned}\tag{155}$$

$$\dot{\le}\exp\Big(-N\min\big\{\gamma,\min_{i\in[M]\setminus\{1\}}\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P_1\right)\le\gamma+\varepsilon_{i,N}}\mathrm{D}\big(V\|P_i\big)\big\}\Big) \tag{156}$$

□

Using Lemmas 14 and 15, we can characterize the achievable error exponent of $\Phi_{\mathrm{seq}}^{(M)}(\gamma)$ as follows

$$\begin{aligned}
&\mathsf{e}_i\left(\Phi_{\mathrm{seq}}^{(M)}(\gamma)\right)=\liminf_{N\to\infty}\frac{-\log\mathrm{P}_i^{\mathrm{err}}(\Phi_{\mathrm{seq}}^{(M)}(\gamma))}{\mathbb{E}_i\left[T_{\mathrm{seq}}^{(M)}\right]}\\
&\ge\min_{j\in[M],j\neq i}\{\theta^{\star}_{i(j),\gamma}\}\times\\
&\quad\liminf_{N\to\infty}\min\left\{\gamma,\min_{j\in[M]\setminus\{i\}}\min_{V:\mathrm{D}\left(V\|P_i\right)\le\gamma+\varepsilon_{j,N}}\mathrm{D}\big(V\|P_j\big)\right\}
\end{aligned}\tag{157}$$

$$\begin{aligned}
&=\min_{j\in[M]\setminus\{i\}}\{\theta^{\star}_{i(j),\gamma}\}\\
&\qquad\times\min\left\{\gamma,\min_{j\in[M]\setminus\{i\}}\min_{V:\mathrm{D}\left(V\|P_i\right)\le\gamma}\mathrm{D}\big(V\|P_j\big)\right\}
\end{aligned}\tag{158}$$

$$=\min_{j\in[M]\setminus\{i\}}\{\theta^{\star}_{i(j),\gamma}\}\gamma \tag{159}$$

$$=\min_{j\in[M]\setminus\{i\}}\mathrm{GJS}\left(P_i,P_j,\theta^{\star}_{i(j),\gamma}\right). \tag{160}$$

where in (157) we use Lemma 15. Then, (158) is obtained using the fact that the optimal value is a continuous function of $\varepsilon_{j,N}$, and $\varepsilon_{j,N}$ converges to zero as $N\to\infty$. We have (159) because

$$\begin{aligned}
&\{\gamma\,|\bigcap_{i=1}^{M}\{\min_{j\in[M],j\neq i}\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P_i\right)\le\gamma}\mathrm{D}\big(V\|P_j\big)\ge\gamma\}\}\\
&=[0,\min_{(i,j)\in\mathcal{M}}C\left(P_i,P_j\right)].
\end{aligned}\tag{161}$$

where $\mathcal{M}\triangleq\{(i,j)\in[M]^2,i\neq j\}$. Finally the last step in (160) follows from (123). Thus, we conclude that the achievable error exponent is obtained as stated in Corollary 3.

Appendix A
Proof of Lemma 8

Lemma 16: Let $\lambda>0$ and let X^N be a sequence drawn from the product distribution Q^N. Also, let $\alpha_N^{\star}$ satisfy the equation $\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^{\star}\right)=\lambda\alpha_N^{\star}$. Here, we assume that the solution exists. Consider the optimization problem

$$\begin{aligned}
U\left(\alpha\right)\;&\triangleq\;\max_{V\in\mathcal{P}(\mathcal{X})}\;\mathrm{GJS}\left(\widehat{Q}_{X^N},V,\alpha\right)\\
&\qquad\text{s.t.}\qquad\mathrm{D}\big(V\|P\big)\le\frac{1}{\sqrt{N}}.
\end{aligned}\tag{162}$$

Then, $\alpha_N^{+}=\alpha_N^{\star}+O\left(\frac{\log N}{N^{\frac{1}{4}}}\right)$ satisfies the following inequality $U\left(\alpha_N^{+}\right)<\lambda\alpha_N^{+}$.

Proof: We begin the proof by rewriting the objective function as

$$\begin{aligned}
&\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha\right)\\
&=\min_{W\in\mathcal{P}(\mathcal{X})}\Big(\sum_{z\in\mathcal{X}}\left(\alpha\widehat{Q}_{X^N}\left(z\right)+V(z)\right)\log 1/W\left(z\right)\Big)\\
&-\alpha\mathrm{H}\left(\widehat{Q}_{X^N}\right)-\mathrm{H}\left(V\right)
\end{aligned}\tag{163}$$

$$\begin{aligned}
&\le-\left(\sum_{z\in\mathcal{X}}\left(\alpha\widehat{Q}_{X^N}\left(z\right)+V\left(z\right)\right)\log\frac{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}{1+\alpha}\right)\\
&+\alpha\sum_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)\log\widehat{Q}_{X^N}\left(z\right)+\sum_{z\in\mathcal{X}}V\left(z\right)\log V\left(z\right)
\end{aligned}\tag{164}$$

where the first step is obtained by using Lemma 2 where we show that GJS can be written in the form of an optimization problem. Setting $W=P$ in the second step, we find an upper bound on the objective function of (162). Then, we obtain

$$\begin{aligned}
&\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha\right)\le\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha\right)\\
&+\mathrm{D}\big(V\|\frac{\alpha\widehat{Q}_{X^N}+P}{1+\alpha}\big)-\mathrm{D}\big(P\|\frac{\alpha\widehat{Q}_{X^N}+P}{1+\alpha}\big)
\end{aligned}\tag{165}$$

$$\begin{aligned}
&\le\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^{\star}\right)+\mathrm{D}\big(\widehat{Q}_{X^N}\|\frac{\alpha_N^{\star}\widehat{Q}_{X^N}+P}{1+\alpha_N^{\star}}\big)\left(\alpha-\alpha_N^{\star}\right)\\
&+\sum_{z\in\mathcal{X}}\left(P\left(z\right)-V\left(z\right)\right)\log\frac{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}{1+\alpha}+\mathrm{H}\left(P\right)-\mathrm{H}\left(V\right)
\end{aligned}\tag{166}$$

$$\begin{aligned}
&\le\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^{\star}\right)+\mathrm{D}\big(\widehat{Q}_{X^N}\|\frac{\alpha_N^{\star}\widehat{Q}_{X^N}+P}{1+\alpha_N^{\star}}\big)\left(\alpha-\alpha_N^{\star}\right)\\
&+\sum_{z\in\mathcal{X}}\left(P\left(z\right)-V\left(z\right)\right)\log\frac{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}{1+\alpha}\\
&-\|P-V\|_1\log\frac{\|P_2-V\|_1}{|\mathcal{X}|}.
\end{aligned}\tag{167}$$

Equation (165) follows from the definitions of GJS and KL divergences. Step (166) comes from the fact that GJS is a concave function in α as shown in Lemma 4. Finally, in the last step we have used [18, Thm. 17.3.3]. Therefore, we have

$$\begin{aligned}
&\max_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P\right)\le\frac{1}{\sqrt{N}}}\mathrm{GJS}\left(\widehat{Q}_{X^N},V,\alpha\right)\\
&\le\max_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\left(V\|P\right)\le\frac{1}{\sqrt{N}}}\sum_{z\in\mathcal{X}}\left(P\left(z\right)-V\left(z\right)\right)\log\frac{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}{1+\alpha}\\
&+\frac{\sqrt{2}}{N^{\frac{1}{4}}}\log\left(\frac{|\mathcal{X}|N^{\frac{1}{4}}}{\sqrt{2}}\right)+\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^{\star}\right)
\end{aligned}$$

$$+\,\mathrm{D}\big(\widehat{Q}_{X^N}\|\frac{\alpha_N^\star\widehat{Q}_{X^N}+P}{1+\alpha_N^\star}\big)\left(\alpha-\alpha_N^\star\right), \tag{168}$$

where (168) is because Pinsker's inequality [18, Lemma 11.6.1] and $\mathrm{D}\big(V\|P\big)\leq\frac{1}{\sqrt{N}}$. Considering the optimization problem in the RHS of (168), we need to provide an upperbound for

$$\begin{aligned}\max_{V\in\mathcal{P}(\mathcal{X})}\quad & \sum_{z\in\mathcal{X}}\left(P\left(z\right)-V\left(z\right)\right)\log\frac{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}{1+\alpha}\\ \text{s.t.}\quad & \|V-P\|_1\leq\frac{\sqrt{2}}{N^{\frac{1}{4}}},\end{aligned} \tag{169}$$

Let us define $\epsilon_z\triangleq V\left(z\right)-P\left(z\right)$ for all $z\in\mathcal{X}$. We can rewrite the optimization problem in (169) as

$$\max_{\epsilon:\sum_{z\in\mathcal{X}}\epsilon_z=0}\quad -\sum_{z\in\mathcal{X}}\epsilon_z\log\frac{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}{1+\alpha} \tag{170a}$$

$$\text{s.t.}\quad\sum_{z\in\mathcal{X}}|\epsilon_z|\leq\frac{\sqrt{2}}{N^{\frac{1}{4}}} \tag{170b}$$

$$-P\left(z\right)\leq\epsilon_z\leq1-P\left(z\right)\quad\forall z\in\mathcal{X} \tag{170c}$$

Because $\min_{z\in\mathcal{Z}}P\left(z\right)>0$, as N becomes large, it is straightforward to verify that the constraints in (170c) will not hold with equality at the optimal point since if so, this would contradict (170b). Thus, we can omit the constraint in (170c). With this simplification, the optimization problem in (170) is in the form of that in Lemma 7, and the optimal value is given by

$$\frac{\sqrt{2}}{N^{\frac{1}{4}}}\log\frac{\max_{z\in\mathcal{X}}\{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)\}}{\min_{z\in\mathcal{X}}\{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)\}}. \tag{171}$$

We can further upper bound the optimal value as

$$\frac{\sqrt{2}}{N^{\frac{1}{4}}}\log\frac{\max_{z\in\mathcal{X}}\{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)\}}{\min_{z\in\mathcal{X}}\{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)\}}\leq$$

$$\frac{\sqrt{2}}{N^{\frac{1}{4}}}\log\left(\frac{\max_{z\in\mathcal{X}}P\left(z\right)}{\min_{z\in\mathcal{X}}P\left(z\right)}\right)+\frac{\sqrt{2}}{N^{\frac{1}{4}}}\frac{\max_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)}{\max_{z\in\mathcal{Z}}P\left(z\right)}\alpha$$

Therefore, plugging (172) into (168) we can provide an upper bound for the optimal value of (162). Finally, letting the upper bound be less than $\lambda\alpha$, we obtain the desired result. □

Lemma 17: Let $\alpha_N^\star$ as defined in Lemma 16.

$$\begin{aligned}U\left(\alpha\right)\ \triangleq\ \min_{V\in\mathcal{P}(\mathcal{X})}\quad & \mathrm{GJS}\left(\widehat{Q}_{X^N},V,\alpha\right)\\ \text{s.t.}\quad & \mathrm{D}\big(V\|P\big)\leq\frac{1}{\sqrt{N}}.\end{aligned} \tag{172}$$

Then, we have $U\left(\alpha_N^-\right)\geq\lambda\alpha_N^-$, where

$$\alpha_N^-=\alpha_N^\star-O\left(\frac{1}{N^{\frac{1}{4}}}\right) \tag{173}$$

Proof: In Lemma 4, we proved that GJS is a convex function in its second argument. Therefore, we can write

$$\begin{aligned}&\mathrm{GJS}\left(\widehat{Q}_{X^N},V,\alpha\right)\geq\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha\right)\\ &\quad+\sum_{z\in\mathcal{X}}\log\frac{(1+\alpha)P\left(z\right)}{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}\left(V\left(z\right)-P\left(z\right)\right),\end{aligned} \tag{174}$$

where we have used the fact that for a convex function f, we have $f(x)\geq f(y)+\nabla f(y)^T(x-y)$ for all x and y. Plugging (174) into (172), we arrive at the following optimization problem:

$$\begin{aligned}\min_{V\in\mathcal{P}(\mathcal{X})}\quad & \sum_{z\in\mathcal{X}}\log\frac{(1+\alpha)P\left(z\right)}{\alpha\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}\left(V\left(z\right)-P\left(z\right)\right)\\ \text{s.t.}\quad & \|V-P\|_1\leq\frac{\sqrt{2}}{N^{\frac{1}{4}}}.\end{aligned} \tag{175}$$

Here, in (175) we have Pinsker's inequality [18, Lemma 11.6.1] in the first constraint. Using Lemma 7, it directly follows that the optimal value of the optimization problem (175) is

$$\frac{1}{\sqrt{2}N^{\frac{1}{4}}}\log\left(\frac{\min_{z\in\mathcal{X}}\{\frac{P(z)}{\alpha\widehat{Q}_{X^N}(z)+P(z)}\}}{\max_{z\in\mathcal{X}}\{\frac{P(z)}{\alpha\widehat{Q}_{X^N}(z)+P(z)}\}}\right) \tag{176}$$

which is straightforward to show that the optimal value can be lower bounded by

$$\frac{1}{\sqrt{2}N^{\frac{1}{4}}}\log\frac{\min_{z\in\mathcal{X}}P\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}-\frac{\alpha}{\sqrt{2}N^{\frac{1}{4}}}\frac{\max_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}. \tag{177}$$

Plugging (177) into (174), we obtain

$$\min_{V\in\mathcal{P}(\mathcal{X}):\mathrm{D}\big(V\|P\big)\leq\frac{1}{\sqrt{N}}}\quad\mathrm{GJS}\left(\widehat{Q}_{X^N},V,\alpha\right) \tag{178}$$

$$\begin{aligned}&\geq\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha\right)+\frac{1}{\sqrt{2}N^{\frac{1}{4}}}\log\frac{\min_{z\in\mathcal{X}}P\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}\\ &\quad-\frac{\alpha}{\sqrt{2}N^{\frac{1}{4}}}\frac{\max_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}.\end{aligned} \tag{179}$$

Fix $0\leq\theta\leq\alpha_N^\star$. From Taylor's theorem, there exists an $\tilde{\theta}\in(\alpha_N^\star-\theta,\alpha_N^\star)$ such that

$$\begin{aligned}&\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^\star-\theta\right)\\ &=\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^\star\right)-\mathrm{D}\big(\widehat{Q}_{X^N}\|\frac{\alpha_N^\star\widehat{Q}_{X^N}+P}{1+\alpha_N^\star}\big)\theta\\ &\quad+\frac{\theta^2}{2(1+\tilde{\theta})}\sum_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)\frac{P\left(z\right)-\widehat{Q}_{X^N}\left(z\right)}{\tilde{\theta}\widehat{Q}_{X^N}\left(z\right)+P\left(z\right)}\end{aligned} \tag{180}$$

$$\begin{aligned}&\geq\mathrm{GJS}\left(\widehat{Q}_{X^N},P,\alpha_N^\star\right)-\mathrm{D}\big(\widehat{Q}_{X^N}\|\frac{\alpha_N^\star\widehat{Q}_{X^N}+P}{1+\alpha_N^\star}\big)\theta\\ &\quad+\frac{\theta^2}{2}\left(\frac{1}{1+\alpha_N^\star}-\sum_{z\in\mathcal{X}}\frac{\widehat{Q}_{X^N}\left(z\right)^2}{P\left(z\right)}\right).\end{aligned} \tag{181}$$

Here, the final step follows by lower bounding the second derivative term. Finally letting the lower bound in (179) be smaller $\lambda\left(\alpha-\theta\right)$, we need to find θ such that

$$\begin{aligned}&\frac{\theta^2}{2}\left(\frac{1}{1+\alpha_N^\star}-\sum_{z\in\mathcal{X}}\frac{\widehat{Q}_{X^N}\left(z\right)^2}{P\left(z\right)}\right)+\theta\lambda\\ &\theta\big(-\mathrm{D}\big(\widehat{Q}_{X^N}\|\frac{\alpha_N^\star\widehat{Q}_{X^N}+P}{1+\alpha_N^\star}\big)+\frac{1}{\sqrt{2}N^{\frac{1}{4}}}\frac{\max_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}\big)\\ &+\frac{1}{\sqrt{2}N^{\frac{1}{4}}}\log\frac{\min_{z\in\mathcal{X}}P\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}-\frac{\alpha_N^\star}{\sqrt{2}N^{\frac{1}{4}}}\frac{\max_{z\in\mathcal{X}}\widehat{Q}_{X^N}\left(z\right)}{\max_{z\in\mathcal{X}}P\left(z\right)}=0\end{aligned} \tag{182}$$

Finally, considering (182) is a quadratic equation in θ and $\alpha_N^- = \alpha_N^\star - \theta$, we obtain the desired result. □

REFERENCES

[1] M. Haghifam, V. Y. F. Tant, and A. Khisti, "Sequential classification with empirically observed statistics," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.

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

[3] L. Zhou, V. Y. F. Tan, and M. Motani, "Second-order asymptotically optimal statistical classification," *Inf. Inference A, J. IMA*, vol. 9, pp. 81–111, Jan. 2019.

[4] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. IT-34, no. 2, pp. 278–286, Mar. 1988.

[5] J. Unnikrishnan and F. M. Naini, "De-anonymizing private data by matching statistics," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2013, pp. 1616–1623.

[6] 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, Jan. 2015.

[7] J. Unnikrishnan and D. Huang, "Weak convergence analysis of asymptotically optimal hypothesis tests," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4285–4299, Jul. 2016.

[8] B. G. Kelly, A. B. Wagner, T. Tularak, and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 782–795, Feb. 2013.

[9] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. Suresh, "Competitive classification and closeness testing," in *Proc. 25th Annu. Conf. Learn. Theory*, Edinburgh, U.K., Jun. 2012, pp. 1–22.

[10] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," *SIAM J. Comput.*, vol. 46, no. 1, pp. 429–455, Jan. 2017.

[11] A. Orlitsky, A. T. Suresh, and Y. Wu, "Optimal prediction of the number of unseen species," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 47, pp. 13283–13288, Nov. 2016.

[12] J. Zou *et al.*, "Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects," *Nature Commun.*, vol. 7, no. 1, p. 13293, Dec. 2016.

[13] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 1014–1019, Sep. 1989.

[14] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh, "A unified maximum likelihood approach for optimal distribution property estimation," 2016, *arXiv:1611.02960*. [Online]. Available: http://arxiv.org/abs/1611.02960

[15] H. He, L. Zhou, and V. Y. F. Tan, "Distributed detection with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4349–4367, Jul. 2020.

[16] H.-W. Hsu and I.-H. Wang, "On binary statistical classification from mismatched empirically observed statistics," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2533–2538.

[17] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, Jun. 1945.

[18] M. T. Cover and A. J. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[19] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[20] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," in *MIT (6.441), UIUC (ECE 563), Yale (STAT 664)*. 2017. [Online]. Available: http://www.stat.yale.edu/~yw562/teaching/itlectures.pdf

[21] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2909–2913, Jan. 2016.

[22] S.-H. Lee, V. Y. F. Tan, and A. Khisti, "Streaming data transmission in the moderate deviations and central limit regimes," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6816–6830, Dec. 2016.

[23] M. Hayashi and V. Y. F. Tan, "Asymmetric evaluations of erasure and undetected error probabilities," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6560–6577, Dec. 2015.

[24] M. Spivak, *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Boca Raton, FL, USA: CRC Press, 2018.

[25] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, vol. 6. Belmont, MA, USA: Athena Scientific, 1997.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[27] R. Durrett, *Probability: Theory and Examples*, vol. 49. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[28] P. Billingsley, *Probability and Measure*. Hoboken, NJ, USA: Wiley, 2008.

Mahdi Haghifam (Member, IEEE) was born in Iran, in 1992. He received the B.Sc. and M.Sc. degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, Canada. His research interests include different aspects of machine learning and information theory, specially applications of the latter in the former.

Vincent Y. F. Tan (Senior Member, IEEE) was born in Singapore, in 1981. He received the B.A. and M.Eng. degrees in electrical and information sciences from Cambridge University in 2005, and the Ph.D. degree in electrical engineering and computer science (EECS) from the Massachusetts Institute of Technology (MIT), in 2011.

He is currently a Dean's Chair Associate Professor with the Department of Electrical and Computer Engineering and the Department of Mathematics, National University of Singapore (NUS). His research interests include information theory, machine learning, and statistical signal processing. He was also an IEEE Information Theory Society Distinguished Lecturer from January 2018 to December 2019. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Associate Editor of machine learning for the IEEE TRANSACTIONS ON INFORMATION THEORY. He is a member of the IEEE Information Theory Society Board of Governors.

Ashish Khisti (Member, IEEE) received the B.A.Sc. degree from the Engineering Science Program, University of Toronto, in 2002, and the master's and Ph.D. degrees from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2004 and 2008, respectively. Since 2009, he has been on the Faculty in the Electrical and Computer Engineering (ECE) Department, University of Toronto, where he was an Assistant Professor from 2009 to 2015, an Associate Professor from 2015 to 2019, and is currently a Full Professor. He also holds the Canada Research Chair of information theory with the ECE Department. His current research interests include theory and applications of machine learning and communication networks. He is also interested in interdisciplinary research involving engineering and healthcare