Achievable Error Exponents for Almost Fixed-Length Hypothesis Testing and Classification

Lin Zhou, Jun Diao and Lin Bai

Abstract

We revisit multiple hypothesis testing and propose a two-phase test, where each phase is a fixed-length test and the second-phase proceeds only if a reject option is decided in the first phase. We derive achievable error exponents of error probabilities under each hypothesis and show that our two-phase test bridges over fixed-length and sequential tests in the similar spirit of Lalitha and Javidi (ISIT, 2016) for binary hypothesis testing. Specifically, our test could achieve the performance close to a sequential test with the asymptotic complexity of a fixed-length test and such test is named the almost fixed-length test. Motivated by practical applications where the generating distribution under each hypothesis is unknown, we generalize our results to the statistical classification framework of Gutman (TIT, 1989). We first consider binary classification and then generalize our results to $M$-ary classification. For both cases, we propose a two-phase test, derive achievable error exponents and demonstrate that our two-phase test bridges over fixed-length and sequential tests. In particular, for $M$-ary classification, no final reject option is required to achieve the same exponent as the sequential test of Haghifam, Tan, and Khisti (TIT, 2021). Our results generalize the design and analysis of the almost fixed-length test for binary hypothesis testing to broader and more practical families of $M$-ary hypothesis testing and statistical classification.

Index Terms

Hypothesis Testing, Classification, Large deviations, Two-phase test, Neyman-Pearson, Bayesian

I. INTRODUCTION

Binary hypothesis testing lies in the intersection of information theory and statistics [1] with applications in various domains. In this problem, one is given two distributions $P_1$ and $P_2$ and a test sequence $Y^n$. The task is to decide from which distribution the test sequence $Y^n$ is generated i.i.d. from and the performance criterion is the type-I and type-II error probabilities. Neyman-Pearson lemma [2] states that the likelihood ratio test (LRT) is optimal. When the type-I error probability is upper bounded by a constant, the Chernoff-Stein lemma [3] shows that the type-II error probability decays exponentially fast at the speed of $D(P_1||P_2)$. Blahut [4] showed that to ensure that the type-I error probability decays exponentially fast with a speed $\lambda$, the decay rate of the type-II error probability is reduced from $D(P_1||P_2)$ to $\min_{\tilde{P}, D(\tilde{P}||P_1) \leq \lambda} D(\tilde{P}||P_2)$. Such a tradeoff can be improved in the sequential setting where the length of the test sequence is allowed to vary but the expected length is upper bounded by $n$. In this setting, Wald [5] showed that the sequential probability ratio test (SPRT) is optimal and achieves the maximal error exponents for both types of error probabilities simultaneously.

The superior performance of the SPRT comes at the high complexity of the test design, where at each time, one takes a new test sample and determines whether to continue collecting samples or to make a decision. One might wonder whether it is possible to achieve the performance of the SPRT while maintaining the simple design and low complexity of a fixed-length test like the LRT. Lalitha and Javidi [6] answered this question affirmatively by proposing an almost fixed-length test and showing that the test could achieve performance close to the SPRT with proper choices of test parameters. Specifically, the almost fixed-length test in [6] consists of two phases: the first phase is a fixed-length test taking $n$ samples, the second phase is another fixed-length test taking additional $\lceil (k-1)n \rceil$ samples for some real number $k \geq 1$ and the second phase proceeds only if the first phase outputs a reject option, indicating that $n$ samples are not sufficient to make a reliable decision.

Motivated by practical decision problems with more than two outcomes, binary hypothesis testing is naturally generalized to the $M$-ary case with more than two hypotheses. In $M$-ary hypothesis testing [7], one is given $M$ probability distributions and a test sequence $Y^n$. The task is to decide from which distribution the test sequence is generated i.i.d. from and the performance criteria is the error probabilities under each hypothesis. For fixed-length tests, Tuncel [8] characterized optimal exponent region. For sequential tests, Baum and Veeravalli [9] proposed the multi-hypothesis sequential probability ratio test (MSPRT) and proved its optimality under the Bayesian setting. Draglia, Tartakovsky and Veeravalli [10] proved that MSPRT is asymptotically optimal not only for the expected sample size but also for any positive moment of the stopping time.

The above results for hypothesis testing are very insightful. However, these results cannot be directly used to guide the test design in practical applications such as image classification, text classification or junk mail identification. This is because the generating distributions, e.g., $P_1$ and $P_2$ in binary hypothesis testing, are usually unknown in these applications. To tackle this problem, Gutman [11] proposed the statistical classification framework where one determines whether a test sequence $Y^n$ is generated from one of the multiple possible unknown distributions using empirically observed statistics. Specifically,

The authors are with the School of Cyber Science and Technology, Beihang University, Beijing, China, 100083.

1 In the rest of the paper, for simplicity, we drop the integer constraint and use $(k-1)n$ and $kn$. 
for the binary case, given two training sequences \(X_1^n\) and \(X_2^n\) that are generated i.i.d. from unknown distributions \(P_1\) and \(P_2\) respectively, one is asked to determine whether the test sequence \(Y^n\) is generated from the same distribution as \(X_1^n\) or \(X_2^n\). Gutman derived the tradeoff of the exponents of type-I and type-II error probabilities. Recently, Haghifam, Tan and Khisti [12, Section II] generalized Gutman’s result to the sequential setting and proved that a sequential test has significantly larger Bayesian error exponent, which is the decay rate of the weighted sum of type-I and type-II error probabilities. The above results were also derived for \(M\)-ary classification.

Inspired by the results of Lalitha and Javidi [6] for binary hypothesis testing, we are interested in the following two questions. Firstly, can one propose a two-phase test for \(M\)-ary hypothesis testing to account for multiple decision outcomes and demonstrate that the test has superior performance close to the sequential test at the asymptotic complexity of a fixed-length test? Secondly, can one generalize the results for hypothesis testing to more practical statistical classification and propose simple two-phase tests for both binary and \(M\)-ary classification? Fortunately, we answer both questions affirmatively. Our contributions are summarized as follows.

A. Main Contributions

Firstly, we generalize the results of Lalitha and Javidi [6] for binary hypothesis testing by allowing more than two hypotheses, proposing a two-phase KL-divergence based test using the empirical distribution of the observed sequence and deriving achievable error exponents of our test. Analytical results and numerical simulations show that by tuning the design parameters, the achievable error exponents of our test approach either the fixed-length test [8] or the sequential test [9] in both Neyman-Pearson and Bayesian settings. The non-asymptotic complexity of our test is at most multiple of a fixed-length test and the asymptotic complexity of our test is the same as the fixed-length test. We emphasize that our test is different from the LRT-based test of Lalitha and Javidi [6, Eq. (14) and Eq. (15)] when specialized to \(M = 2\), but it is more amenable to generalizations to the case with unknown distributions.

Secondly, we generalize the results for hypothesis testing to the more practical statistical classification problem where the generating distribution under each hypothesis is unknown. We consider both binary and \(M\)-ary classification. For both cases, we propose two-phase tests based on the generalized Jensen-Shannon divergence [13, Eq. (2.3)] that use empirical distributions of training sequences and the testing sequence and derive achievable error exponents of our tests. Furthermore, we further discuss how the performance of our tests are influenced by the ratio of the lengths of training sequences and that of the test sequence. Analytical results and numerical simulations verify that our tests bridge the performance gaps between the fixed-length test [11] and the sequential test [12, Eq. (7) and Eq. (24)] for statistical classification by having performance close to the sequential test with the asymptotic complexity of a fixed-length test. In particular, for \(M\)-ary classification, our two-phase test does not require an additional reject option to achieve the Bayesian error exponent of the sequential test in [12, Eq. (24)].

B. Other Related Works

We provide a non-exhaustive list of related works on statistical classification. Zhou, Tan and Motani [13] analyzed the finite sample performance of Gutman’s test. Hsu and Wang [14] generalized Gutman’s results to the mismatched setting where the generating distributions of the training sequences and the test sequences deviate slightly. The Bayesian error probability of binary classification was studied by Merhav and Ziv [15] in the asymptotic case and more recently by Saito [16] in the finite blocklength using Bayes codes. Gutman’s results are also generalized to multiple test sequences [17], distributed detection [18], quickest change-point detection [19], outlier hypothesis testing [20], [21] and universal sequential classification [22].

Notation

We use \(\mathbb{R}, \mathbb{R}_+, \mathbb{N}\) to denote the set of real numbers, non-negative real numbers, and natural numbers respectively. Given any integer \(a \in \mathbb{N}\), we use \([a]\) to denote \([1, 2, \ldots, a]\). Random variables and their realizations are denoted by upper case variables (e.g., \(X\)) and lower case variables (e.g., \(x\)), respectively. All sets are denoted in calligraphic font (e.g., \(\mathcal{X}\)). For any \(N \in \mathbb{N}\), let \(X^N := (X_1, \ldots, X_N)\) be a random vector of length \(N\) and let \(x^N = (x_1, \ldots, x_N)\) be a particular realization of \(X^N\). The set of all probability distributions on a finite set \(\mathcal{X}\) is denoted as \(\mathcal{P}(\mathcal{X})\). We use \(\mathbb{E}[\cdot]\) to denote expectation. Given a sequence \(x^n \in \mathcal{X}^n\), the type or empirical distribution is defined as \(\hat{T}_{x^n}(a) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i) = a, \forall a \in \mathcal{X}\). The set of types formed from length-\(n\) sequences with alphabet \(\mathcal{X}\) is denoted as \(\mathcal{P}^n(\mathcal{X})\). Given any \(P \in \mathcal{P}^n(\mathcal{X})\), the set of all sequences of length \(n\) with type \(P\), the type class, is denoted as \(T_P^n\).

II. Almost Fixed-length \(M\)-ary Hypothesis Testing

A. Problem formulation

Consider a finite alphabet \(\mathcal{X}\). In multiple hypothesis testing, we are given an observed sequence \(Y^\tau = (Y_1, \ldots, Y_\tau) \in \mathcal{X}^\tau\) where \(\tau\) is a random stopping time with respect to the filtration \(\sigma\{Y_1, \ldots, Y_n\}\). The observed sequence is generated i.i.d. from one of \(M\) distributions \(P := (P_1, \ldots, P_M) \in \mathcal{P}(\mathcal{X})^M\) and thus there exists \(M\) hypotheses \(\{H_j\}_{j \in [M]}\) which correspond to \(M\) possible underlying known distributions. The task for multiple hypothesis testing problem is to design a test \(\Phi = (\tau, \phi_\tau)\),
which consists of a random stopping time \( \tau \) and a mapping \( \phi_\tau : \mathcal{X}^\tau \rightarrow \{H_1, \ldots, H_M\} \), to decide among the following \( M \) hypotheses:

- \( H_j \): the sequence \( Y^\tau \) is generated i.i.d. from the distribution \( P_j, j \in [M] \).

At the random stopping time \( \tau \), any mapping \( \phi_\tau \) partitions the sample space into \( M \) disjoint regions: \( \{A_j(\phi_\tau)\}_{j \in [M]} \) where \( Y^\tau \in A_j(\phi_\tau) \) favors hypothesis \( H_j \).

Given any test \( \Phi = (\tau, \phi_\tau) \), for any tuple of distributions \( P \in \mathcal{P}(\mathcal{X})^M \), we evaluate the performance of the test via the following \( M \) error probabilities:

\[
\beta_j(\Phi|P) = P_j\{\phi_\tau(X^\tau) \neq H_j\},
\]

where we define \( P_j\{\cdot\} := \Pr\{\cdot|H_j\} \) under which the test sequence \( Y^\tau \) is generated i.i.d. from the distribution \( P_j \). The quantity \( \beta_j(\Phi|P) \) is known as the type-\( j \) error probability.

In certain cases, one is also interested in the Bayesian error probability. Assume that prior probabilities \( \pi_1, \ldots, \pi_M \) for hypotheses \( H_1, \ldots, H_M \) respectively, such that \( \pi_j \in (0,1) \) for each \( j \in [M] \) and \( \sum_{j=1}^M \pi_j = 1 \). The Bayesian error probability is the sum of weighted error probabilities defined as follows:

\[
P_{\text{Bayesian}}(\Phi|P) = \sum_{j=1}^M \pi_j \beta_j(\Phi|P).
\]

It is challenging to characterize the non-asymptotic performance of error probabilities with finite sample size. As a compromise, consistent with literature, one usually derives error exponents, i.e., the decay rates of error probabilities. For \( M \)-ary hypothesis testing, we are interested in the following error exponents

\[
E_j(\Phi|P) := \liminf_{n \to \infty} \frac{-\log \beta_j(\Phi|P)}{E_n[\tau]}, \quad j \in [M],
\]

and the Bayesian error exponent

\[
E_{\text{Bayesian}}(\Phi|P) := \liminf_{n \to \infty} \frac{-\log P_{\text{Bayesian}}(\Phi|P)}{E_n[\tau]}.
\]

Note that in both (3) and (4), the random stopping time \( \tau \) is a function of \( n \) through the filtration \( \sigma\{Y_1, \ldots, Y_n\} \) and thus the definitions are valid.

Fix any integer \( n \in \mathbb{N} \). When the random stopping time \( \tau \) is fixed as a constant such that \( \tau = n \), the test is called a fixed-length test; when the random stopping time \( \tau \) satisfies \( E_n[\tau] \leq n \), the test is called a sequential test. Optimal error exponent results for both tests were characterized in [8] and [9], respectively, which we recall in the next subsection.

### B. Existing results

We first recall error exponents for fixed-length \( M \)-ary hypothesis testing by Tuncel [8]. The exponents \( E := (E_1, \ldots, E_M) \) are called achievable if there exists a sequence of tests \( \Phi = (n, \phi_n) \) satisfying

\[
E_j(\Phi|P) \geq E_j, \quad j \in [M].
\]

The set of all achievable error exponents is called the error exponent region and denoted by \( \mathcal{R} \).

Given any \( j \in [M] \), define the set \( \mathcal{M}_j := \{i \in [M] : i \neq j\} \). Tuncel [8] proved the following result.

**Theorem 1.** Given any tuple of distributions \( P \in \mathcal{P}(\mathcal{X})^M \), the error exponent region \( \mathcal{R} \) satisfies:

\[
\mathcal{R} = \{E = (E_1, \ldots, E_M) : E_j = \min_{i \in \mathcal{M}_j} T_{j,i}, \forall j \in [M] \text{ for some } T := \{T_{j,i}\}_{j \neq i} \in \mathcal{R}_{\text{tuncel}}\},
\]

where

\[
\mathcal{R}_{\text{tuncel}} = \{T : \forall P \in \mathcal{P}(\mathcal{X}) \exists i \in [M] \text{ s.t. } T_{j,i} \leq D(P||P_j) \text{ for all } j \neq i\}.
\]

We then recall error exponents for sequential tests by Baum and Veeravalli [9]. For each \( j \in [M] \), let \( \pi_j \) denote the prior probability of hypothesis \( H_j \) and \( \tilde{h}_j^n \) denote the following posterior probability with \( n \) test samples \( Y^n \):

\[
\tilde{h}_j(Y^n) = \frac{\pi_j \prod_{i=1}^n P_j(Y_i)}{\sum_{k=1}^M \pi_k \left(\prod_{i=1}^n P_k(Y_i)\right)}.
\]

Given a threshold \( T > 1 \), the sequential test \( \Phi_{\text{MST}} = (\tau_{\text{MST}}, \phi_{\tau_{\text{MST}}}) \) consists of the random stopping time

\[
\tau_{\text{MST}} = \inf \left\{n \in \mathbb{N} : \exists j \in [M] \text{ such that } \tilde{h}_j(Y^n) > T \right\},
\]
and the decision rule
\[
\phi_{MST} = H_j, \quad j = \arg\max_{i \in [M]} h_i(Y^{MST}). \tag{10}
\]

Baum and Veeravalli \cite{BaumVeeravalli} proved the following result.

**Theorem 2.** Given any tuple of distributions \( P \in \mathcal{P}(X)^M \), the test \( \Phi_{MST} \) achieves the following error exponents
\[
E_j(\Phi|P) \geq \min_{i \in [M]} D(P_i\|P_j), \quad j \in [M]. \tag{11}
\]

Conversely, the test \( \Phi_{MST} \) is asymptotically optimal in the sense that among all tests \( \Phi = (\tau, \phi_\tau) \) achieving the same error exponents, \( \Phi_{MST} \) has smaller average sample size, i.e., \( \mathbb{E}_j[\tau] \geq \mathbb{E}_j[\tau_{MST}] \) for all \( j \in [M] \).

Note that Theorem 1 implies that for each \( j \in [M] \), \( \min_{i \in [M]} D(P_i\|P_j) \) is the maximal achievable exponent of the type-\( j \) error probability of a fixed-length test at the cost of smaller exponents for other error probabilities. In contrast, Theorem 2 implies that the optimal sequential test achieves the maximal error exponents simultaneously for all \( M \) hypotheses. However, it is critical to note that the sequential test achieves the superior performance at the cost of very high computation complexity since one needs to determine whether to stop after collecting each new test sample, which renders sequential tests impractical for resource-limited applications.

One might wonder whether it is possible to achieve performance close to the optimal sequential test at a complexity close to a fixed-length test. In the next subsection, we answer this question affirmatively by proposing a two-phase test, analyzing its achievable error exponents and showing that its achievable error exponents approach the optimal sequential test \( \Phi_{MST} \) with the same asymptotic complexity as a fixed-length test, numerically and analytically.

**C. Our Two-phase Test**

In this subsection, we present our two-phase test that uses empirical distributions of the testing sequences and proceeds in two phases. In the first phase, the test takes \( n \) samples \( Y^n \) to perform a fixed-length test with a reject option, where reject means that more samples are needed to make a reliable decision. Once a reject decision is made in the first phase, our test proceeds in the second phase, where \((k-1)n\) additional test samples \( Y_{k+1}^{kn} \) are collected and a fixed-length test without a rejection is used to make a final decision.

To present our test, we need the following notation. Given any integers \( (M,n) \in \mathbb{N}^2 \) and any tuple of distributions \( P \in \mathcal{P}(X)^M \), define the set \( \mathcal{N} := [M] \setminus \arg \min_{i \in [M]} D(\hat{T}_{Y^n}\|P_i) \). Fix any positive real numbers \( (k, \lambda_1, \ldots, \lambda_M) \in \mathbb{R}_{+}^{M+1} \). The random stopping time \( \tau \) of our test is either \( n \) or \( kn \) depending on the first \( n \) test sequences \( Y^n \), i.e.,
\[
\tau = \begin{cases} 
n & \text{if } \forall i \in \mathcal{N} \ D(\hat{T}_{Y^n}\|P_i) > \lambda_i, \\
k n & \text{otherwise.} \end{cases} \tag{12}
\]

The decision rule \( \phi_\tau \) of our test applies nearest neighbor (NN) detection based on KL divergence between the empirical distributions of the testing sequence and the generating distributions. Specifically, when \( \tau = n \), our test \( \phi_n \) favors hypothesis \( H_j \) if the KL divergence \( D(\hat{T}_{Y^n}\|P_j) \) is smallest among all \( \{D(\hat{T}_{Y^n}\|P_i)\}_{i \in [M]} \), i.e.,
\[
\phi_n(Y^n) = H_j, \quad j = \arg \min_{i \in [M]} D(\hat{T}_{Y^n}\|P_i). \tag{13}
\]

When \( \tau = kn \), we use the same test with the only exception that the empirical distribution \( \hat{T}_{Y^{kn}} \) of all \( kn \) testing sequences is used instead of \( \hat{T}_{Y^n} \), i.e.,
\[
\phi_{kn}(Y^{kn}) = H_j, \quad j = \arg \min_{i \in [M]} D(\hat{T}_{Y^{kn}}\|P_i). \tag{14}
\]

For simplicity, we use \( \Phi_{J,t}^{(M)} \) to denote our two-phase test for \( M \)-ary hypothesis testing. Our test belongs to the family of \((\gamma, k)\)-almost fixed-length test \cite{Weinstein1964} that includes any test \( \Phi = (\tau, \phi_\tau) \) such that \( \mathbb{P}_i\{\tau > n\} \leq \exp(-n\gamma) \) and \( \tau \leq kn \) for some \((\gamma, k) \in \mathbb{R}_+^2 \). Asymptotically, such a test ensures that the average stopping time satisfies \( \limsup_{n \to \infty} E[\tau]/n \leq 1 \), which achieves the same asymptotic sample complexity as a fixed-length test.

We then discuss the asymptotic performance of our two-phase test. Intuitively, for each \( j \in [M] \), if the testing sequence \( Y^n \) is generated from the distribution \( P_j \), the empirical distribution of \( \hat{T}_{Y^n} \) tends to \( P_j \) as the length of sequence \( n \) increases. Thus, the KL divergence \( D(\hat{T}_{Y^n}\|P_j) \) tends to zero and all other \( M-1 \) KL divergences \( D(\hat{T}_{Y^n}\|P_i), i \in \mathcal{N} \) tend to the positive number \( D(P_i\|P_j) \). With high probability, a correct decision can be made in the first phase when \( n \) is large. In an exponentially rare case where a reject is decided in the first phase, decision maker can collect extra \((k-1)n\) samples for \( k \) large and make a reliable choice in the second phase with the same logic.
D. Main Results and Discussions

We need the following definitions to present our theoretical results. Define the set
\[ \mathcal{M}_{\text{dis}} = \{(i,j) \in [M]^2 : i \neq j\}. \] (15)

Given any tuple of distributions \( P = (P_1, \ldots, P_M) \in \mathcal{P}(\mathcal{X})^M \) and any thresholds \( \lambda^M = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}_+^M \), define the exponent function
\[ \Gamma_j(\lambda^M | P) = \min_{(i,j) \in \mathcal{M}_{\text{dis}}} \min_{Q \in \mathcal{P}(\mathcal{X})^M} \min_{D(Q || P_j), D(Q || P_i) \leq \lambda_i} D(Q || P_j), \] (16)

As we shall show, \( \Gamma_j(\lambda^M | P) \) is critical to bound the exponential decaying probability that the random stopping time exceeds \( n \), i.e., \( \Pr(\tau > n) \leq \exp(-n\gamma) \) for some targeted exponent \( \gamma \in \mathbb{R}_+ \).

For each \( j \in [M] \), note that \( \Gamma_j(\lambda^M | P) \) is non-increasing in \( \lambda_j \). In particular, \( \Gamma_j(\lambda^M | P) = 0 \) when
\[ \lambda_j \geq \min_{(i,j) \in \mathcal{M}_{\text{dis}}} \max\{D(P_j || P_i), D(P_j || P_i)\}. \] (17)

The maximal finite value of the right hand side of (16) is denoted as \( \bar{\gamma}_j(P) \).

Furthermore, given any \( k \in \mathbb{R}_+ \), define the exponent function
\[ \Omega_j(k | P) = \min_{i \in \mathcal{M}_j} \min_{Q \in \mathcal{P}(\mathcal{X})^M} \min_{D(Q || P_j), D(Q || P_i) < D(Q || P_j)} kD(Q || P_j). \] (18)

As we shall show, \( \Omega_j(k | P) \) characterizes the \( j \)-th error exponent in the second phase of our test with \( kn \) samples.

With these definitions, our result states as follows.

**Theorem 3.** For any \( (k, \gamma) \in \mathbb{R}_+^2 \) and any tuple of distributions \( P \in \mathcal{P}(\mathcal{X})^M \), the achievable type-\( j \) error exponents of our two-phase test satisfies that for each \( j \in [M] \),
\[ E_j(\Phi^{(M)}_{\text{ht}} | P) \geq \min \{ \lambda_j, \Omega_j(k | P) + \gamma \}, \] (19)
where the thresholds \( \lambda^M = (\lambda_1, \ldots, \lambda_M) \) satisfy \( \lambda^M \in \tilde{G}(\gamma) := \{ \lambda^M \in \mathbb{R}_+^M : \min_{j \in [M]} \Gamma_j(\lambda^M | P) \geq \gamma \} \). Furthermore, the Bayesian error exponent of our two-phase test satisfies
\[ E_{\text{Bayesian}}(\Phi^{(M)}_{\text{ht}} | P) \geq \max_{\lambda^M \in \tilde{G}(\gamma)} \min_{j \in [M]} \{ \lambda_j, \Omega_j(k | P) + \gamma \}. \] (20)

The proof of Theorem 3 is given in Appendix A. We make several remarks.

Firstly, our results generalize the results of Lalitha and Javidi [6] for binary hypothesis testing to multiple hypothesis testing with more than two decision outcomes. In the same spirit of [6], our two-phase test bridges over the fixed-length test [8] and the sequential test [9] since the achievable exponents of our test approach exponents of either case with proper choices of parameters \( \gamma \) and \( k \). Specifically, under any tuple of distributions \( P \), if \( \gamma \) is large enough so that \( \gamma > \min_{j \in [M]} \bar{\gamma}_j(P) \), the set \( \tilde{G}(\gamma) \) contains only the all zero vector, and thus the only valid thresholds \( \lambda^M \) of our test are all zero. In this case, our two-phase test reduces to a fixed-length test using \( n \) samples (cf. [12]). On the other hand, if \( \gamma \to 0 \), for each \( j \in [M] \), the maximum threshold \( \lambda_j \) approaches \( \min_{(i,j) \in \mathcal{M}_{\text{dis}}} \max\{D(P_j || P_i), D(P_j || P_i)\} \). With proper choice of \( k \), the achievable exponents of our test approaches the achievable exponent of the optimal sequential test [9]. Finally, if \( \gamma \in (0, \bar{\gamma}(P)) \), our two-phase test has performance in between the fixed-length test and the sequential test. In the same logic, the achievable Bayesian exponent of our test also bridges that of a fixed-length and a sequential test by tuning parameters \( \gamma \) and \( k \).

To illustrate our results, we run numerical examples to plot the achievable exponents of our two-phase test and compare our results with optimal achievable exponents of the fixed-length test in Theorem 1 and the sequential test in Theorem 2. Specifically, in Fig. 1, we plot the achievable type-I and type-II error exponents of two-phase test with various values of \( \gamma \) and \( k \) when \( M = 2 \). By tuning the parameters, when \( \gamma \) is large, the performance of our test approaches the fixed-length test and when \( \gamma \) is small, the performance of our test approaches the sequential test as desired. Furthermore, in Table 1, we numerically compare the achievable Bayesian error exponent of our test for \( M = 3 \) with fixed-length and sequential tests under distributions \( P_1 = [0.3, 0.3, 0.4], P_2 = [0.4, 0.5, 0.1] \) and \( P_3 = [0.1, 0.7, 0.2] \). The numerical results verify that the performance of test bridges over that of fixed-length and sequential tests by tuning the parameters of \( (k, \gamma) \).

Finally, we remark that our two-phase test is based on the Hoeffding’s test [23] and is different from the test by Lalitha and Javidi [6] that is based on the likelihood ratio test. Specifically, the test in [6] compares the log likelihood ratios with different thresholds when \( M = 2 \) while our test uses the KL divergence based nearest neighbor detection rule with thresholds. We find that our test is more amenable to be generalized to statistical classification where the generating distributions are unknown.
Fixed-length hypothesis testing
Sequential hypothesis testing
Almost fixed-length hypothesis testing(γ = 10⁻⁸, k = 10)
Almost fixed-length hypothesis testing(γ = 0.001, k = 2.2)
Almost fixed-length hypothesis testing(γ = 0.005, k = 1.8)
Almost fixed-length hypothesis testing(γ = 0.01, k = 1.1)
Fixed-length hypothesis testing

Fig. 1. Illustration of the type-I and type-II error exponents of multiple hypothesis testing problem for distributions $P_1 = [0.9, 0.1]$ and $P_2 = [0.2, 0.8]$ when $M = 2.$

### TABLE I

**Bayesian Error Exponents of Multiple Hypothesis Testing.**

| Parameters                                      | Bayesian error exponent |
|------------------------------------------------|-------------------------|
| Sequential hypothesis testing                   | 0.2319                  |
| Almost fixed-length hypothesis testing (γ = 10⁻⁸, k = 10) | 0.2310                  |
| Almost fixed-length hypothesis testing (γ = 0.001, k = 2.2) | 0.2013                  |
| Almost fixed-length hypothesis testing (γ = 0.005, k = 1.8) | 0.1647                  |
| Almost fixed-length hypothesis testing (γ = 0.01, k = 1.1) | 0.1007                  |
| Fixed-length hypothesis testing                 | 0.0936                  |

III. Almost Fixed-length Binary Classification

In this section, we generalize our results to statistical classification [11] where the generating distribution under each hypothesis is unknown but training sequences are available. We first consider binary classification for ease of presentation and then generalize our results to multiple hypotheses in the next section.

**A. Problem formulation**

In binary classification, we are given two length $N \in \mathbb{N}$ training sequences $(X_1^N, X_2^N)$ that are generated i.i.d. from unknown distributions $P_1$ and $P_2$ respectively. The test sequence $Y^\tau = (Y_1, \ldots, Y_\tau)$ is generated i.i.d. from either unknown $P_1$ or $P_2$, where $\tau$ is a random stopping time with respect to the Filtration $\sigma \{X_1^N, X_2^N, Y_1, \ldots, Y_n\}$. The task for binary classification is to design a test $\Phi$ ignorant of generating distributions, which consists of a random stopping time $\tau$ and a mapping $\phi_\tau: \mathcal{X}^{2N} \times \mathcal{X}^\tau \to \{H_1, H_2\}$ in order to decide between the following two hypotheses:

- $H_1$: the test sequence $Y^\tau$ is generated i.i.d. from the same distribution as the first training sequence $X_1^N$.
- $H_2$: the test sequence $Y^\tau$ is generated i.i.d. from the same distribution as the second training sequence $X_2^N$.

Note that given a stopping time $\tau$, any mapping $\phi_\tau$ partitions the sample space into two disjoint regions: $\mathcal{A}(\phi_\tau)$ where $(X_1^N, X_2^N, Y^\tau) \in \mathcal{A}(\phi_\tau)$ favors hypothesis $H_1$ and $\mathcal{A}^c(\phi_\tau)$ where $(X_1^N, X_2^N, Y^\tau) \in \mathcal{A}^c(\phi_\tau)$ favors hypothesis $H_2$. For simplicity, we assume that $N = \lceil n\alpha \rceil$.

Given any test $\Phi = (\tau, \phi_\tau)$, for any pair of generating distributions $(P_1, P_2)$, we evaluate the performance of the test via the following type-I and type-II error probabilities:

$$
\beta_1(\Phi|P_1, P_2) = P_1 \{\phi_\tau(X_1^N, X_2^N, Y^\tau) \neq H_1\},
$$

**In subsequent analysis, without loss of generality, we ignore the integer requirement and drop the ceiling operation for simplicity.**
where for each \( i \in [2] \), we define \( \mathbb{P}_i \{ \cdot \} := \Pr \{ \cdot \mid H_i \} \) where \( X_1^N \sim P_1, X_2^N \sim P_2 \) and \( Y^\tau \sim P_i \). Specifically, we are interested in the error exponents of both error probabilities:

\[
E_i(\Phi|P_1,P_2) := \lim \inf_{N \to \infty} \frac{-\log \beta_i(\Phi|P_1,P_2)}{E_i[\tau]}.
\]  

(22)

Similar to hypothesis testing, two constrains of random stopping time \( \tau \), i.e., \( \tau = n \) and \( E_i[\tau] \leq n \), correspond to a fixed-length test and a sequential test, respectively. Error exponent results for these two cases were derived in [11], [12, Theorem 2], which we shall recall in the next subsection.

B. Existing results

We first recall error exponents of a fixed-length test by Gutman [11]. To present Gutman’s test, we need the following measure of distributions. Given any two distributions \( B \). Existing results [2], which we shall recall in the next subsection.

Conversely, Gutman’s test is asymptotically optimal in the generalized Neyman-Pearson sense. Specifically, for all fixed-length tests \( \Phi = (n, \phi_n) \) that ensure \( E_1(\Phi|P_1,P_2) \geq \lambda \) for all pairs of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X})^2 \), given any pair of distributions \((P_1, P_2) \), \( E_2(\Phi|P_1, P_2) \leq E_2(\Phi_{\text{Gut}}|P_1, P_2) \).

We then recall the error exponents of a sequential test by Haghifam et al. [12, Theorem 2]. The sequential test \( \Phi_{\text{seq}} = (\tau_{\text{seq}}, \phi_{\tau_{\text{seq}}}) \) consists of the random stopping time

\[
\tau_{\text{seq}} = \min \left( N^2, \inf \{ n \in \mathbb{N} : \exists \ i \in [2], \ GJS(\tilde{T}_X^n, \tilde{T}_Y^n, N/n) > \mu N \} \right),
\]

(28)

and the decision rule

\[
\phi_{\tau_{\text{seq}}} = \begin{cases} 
H_1 & \text{if } GJS(\tilde{T}_X^n, \tilde{T}_Y^n, \frac{N}{\tau_{\text{seq}}}) \leq \mu N, \\
H_2 & \text{otherwise},
\end{cases}
\]

(29)

where \( \mu \in \mathbb{R}_+ \) is a design parameter.

To present the results in [12, Theorem 2], we need the following definition of Chernoff information for any distributions \((P_1, P_2) \)

\[
C(P_1, P_2) = -\min_{\rho \in [0,1]} \log \sum_{x \in \mathcal{X}} P_1(x)^\rho P_2(x)^{1-\rho}.
\]

(30)

Haghifam et al. proved the following result.

Theorem 5. Consider any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X})^2 \) and any \( \mu \in (0, C(P_1, P_2)] \). Let \( \beta_\mu^* \in \mathbb{R}_+ \) and \( \theta_\mu^* \in \mathbb{R}_+ \) be the solutions of \( GJS(P_2, P_1, \beta_\mu^*) = \mu \beta_\mu^* \) and \( GJS(P_1, P_2, \theta_\mu^*) = \mu \theta_\mu^* \), respectively. Then it follows that

\[
E_1(\Phi_{\text{seq}}|P_1, P_2) \geq GJS(P_2, P_1, \beta_\mu^*),
\]

(31)

\[
E_2(\Phi_{\text{seq}}|P_1, P_2) \geq GJS(P_1, P_2, \theta_\mu^*).
\]

(32)
As proved in [12, Theorem 3] and demonstrated in [12, Fig. 2], the sequential test $\Phi_{seq}$ has larger Bayesian error exponent than the fixed-length test $\Phi_{Gut}$. However, the computation complexity of the sequential test of $\Phi_{seq}$ is much higher for the same reason discussed for sequential hypothesis testing. One might wonder whether it is possible to achieve performance close to $\Phi_{seq}$ at a complexity similar to $\Phi_{Gut}$ by generating the results of Lalitha and Javidi [6] for binary hypothesis testing. In the next section, we answer this question affirmatively by proposing a two-phase test, analyzing its error exponent and demonstrating that our test can achieve error exponent close to $\Phi_{seq}$ with the same asymptotic average test sequence length as $\Phi_{Gut}$ with proper choices of design parameters.

C. Our Two-phase Test

Consider any integer $n \in \mathbb{N}$ and any positive real numbers $(k, \lambda_1, \lambda_2, \lambda) \in \mathbb{R}_+^4$. Recall the definition of the generalized Jensen-Shannon divergence $GJS(\cdot)$ in [23]. Our test is an almost fixed-length test with two phases where the random stopping time $\tau$ is either $n$ or $kn$ depending on training and testing sequences. Specifically, the random stopping time of our test satisfies

$$\tau = \begin{cases} n & \text{if } \exists i \in [2], \ GJS(\hat{T}_{X_i}^N, \hat{T}_{Y_i}^n, \alpha) > \lambda_i, \\ kn & \text{otherwise}, \end{cases}$$

(33)

where $GJS(\hat{T}_{X_i}^N, \hat{T}_{Y_i}^n, \frac{N}{n})$ can be understood as the scoring function. Note that both thresholds $\lambda_1$ and $\lambda_2$ are used to determine the random stopping time $\tau$. Note that $\tau = kn$ implies that the samples collected till time $n$ is not sufficient to make a reliable decision between the two hypotheses and corresponds to a reject option for a fixed-length test with $n$ samples [13]. We then specify the decision rules for the two-phase test. When $\tau = n$, we use the following Gutman’s test with a threshold $\lambda_1$

$$\phi_n(X_1^N, X_2^N, Y^n) = \begin{cases} H_1 & \text{if } GJS(\hat{T}_{X_1}^N, \hat{T}_{Y^n}, \alpha) \leq \lambda_1, \\ H_2 & \text{otherwise}. \end{cases}$$

(34)

When $\tau = kn$, we use another Gutman’s test with a threshold $\lambda$ as follows:

$$\phi_{kn}(X_1^N, X_2^N, Y^{kn}) = \begin{cases} H_1 & \text{if } GJS(\hat{T}_{X_1}^N, \hat{T}_{Y^{kn}}, \frac{n}{\lambda}) \leq \lambda, \\ H_2 & \text{otherwise}. \end{cases}$$

(35)

For simplicity, we use $\Phi_{tp}$ to denote the above two-phase test. If $\lambda_2 = 0$, the stopping time $\tau$ is always $n$ and the two-phase test $\Phi_{tp}$ reduces the Gutman’s fixed-length test $\Phi_{Gut}$. Note that our two-phase test also belongs to the family of $(\gamma, k)$-almost fixed-length test [6] which achieves the same sample complexity with the fixed-length Gutman’s test asymptotically.

We then comment on the asymptotic performance of our two-phase test. Intuitively, for each $i \in [2]$, if the test sequence $Y^n$ is generated from the unknown distribution $P_i$, then as the length of each observed sequence $n$ increases, the empirical distribution of $\hat{T}_{Y^n}$ tends to $P_i$. Similarly, $\hat{T}_{X_1}^N$ and $\hat{T}_{X_2}^N$ tend to $P_1$ and $P_2$ respectively. Thus, for each $i \in [2]$, under hypothesis $H_i$, the scoring function $GJS(\hat{T}_{X_i}^N, \hat{T}_{Y^n}, \alpha)$ tends to zero and the other scoring function $GJS(\hat{T}_{X_j}^N, \hat{T}_{Y^n}, \alpha)$ tends to $GJS(P_j, P_i, \alpha)$ where $j \in [2]$ and $j \neq i$. Thus, in the same logic as hypothesis testing, a reliable decision can be made with high probability.

D. Main Results and Discussions

We need the following definitions to present our results. Given any $\alpha \in \mathbb{R}_+$ and any $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, for each $i \in [2]$, let $i = 3 - i$ and define the following exponent function

$$F_i(\alpha, \lambda_i|P_1, P_2) = \min_{(Q_1, Q_2) \in \mathcal{P}(\mathcal{X})^2} \alpha D(Q_i||P_1) + D(Q_i||P_1).$$

(36)

As we shall show, $F_i(\alpha, \lambda_i|P_1, P_2)$ is critical to bound the probability that our test proceeds the second phase, i.e., $\Pr\{\tau > n\}$. Note that for each $i \in [2]$, $F_i(\alpha, \lambda_i|P_1, P_2)$ decrease in $\lambda_i$. In particular, $F_i(\alpha, \lambda_i|P_1, P_2) = 0$ if $\lambda_i \geq GJS(P_i, P_1, \alpha)$ and achieves the following maximum when $\lambda_i = 0$

$$\min_{Q \in \mathcal{P}(\mathcal{X})} \alpha D(Q||P_1) + D(Q||P_1) = \left(1 + \alpha\right) \log \sum_{x} P_1(x)^{\frac{1}{\alpha+1}} P_i(x)^{\frac{1}{1+\alpha}}$$

(37)

$$= D_{\frac{1}{\alpha+1}}(P_i||P_1),$$

(38)

which is the Rényi divergence of order $\frac{1}{\alpha+1}$ [13, Eq. (3.16)].

Furthermore, given any $k \in \mathbb{R}_+$, let

$$H(k, \alpha, \lambda|P_1, P_2) = \min_{(Q_1, Q_2) \in \mathcal{P}(\mathcal{X})^2} \alpha D(Q_1||P_1) + k D(Q_2||P_2).$$

(39)
As we shall show, $H(k, \alpha, \lambda | P_1, P_2)$ characterizes the type-II error exponent in the second phase of our test where we use a Gutman’s test with $kn$ samples and the threshold $\lambda$.

With these definitions, our main result states as follows.

**Theorem 6.** For any $n \in \mathbb{N}$, $k \in \mathbb{R}_+$ and $\gamma \in \mathbb{R}_+$, under any pair of distributions $(P_1, P_2)$, the achievable type-I and type-II error exponents of our two-phase test satisfy

$$E_1(\Phi_{tp}|P_1, P_2) \geq \min \{ \lambda_1, k\lambda + \gamma \},$$

$$E_2(\Phi_{tp}|P_1, P_2) \geq \min \{ \lambda_2, H(k, \alpha, \lambda | P_1, P_2) + \gamma \},$$

where $(\lambda_1, \lambda_2) \in \mathbb{R}_+$ satisfy $\min_{i\in[2]} F_i(\alpha, \lambda_i|P_1, P_2) \geq \gamma$ and $\lambda_2 \neq 0$. When $\lambda_2 = 0$, the two-phase test reduces to Gutman’s test and thus the achievable error exponents reduce to Gutman’s [7] Theorem 3.

The proof of Theorem 6 is provided in Appendix B. We remark that the proof of Theorem 6 is similar to binary classification with a reject option [13]. The probability that our test enters a second phase equals the probability that a reject option is decided in binary classification, which implies that further investigation is needed. The same probability was studied in the second-order asymptotic regime in [13] while we study the large deviations regime. The derivations of the error exponents in both phases of our test follow Gutman [11].

We also make several other remarks. Firstly, Theorem 6 generalizes the results of Lalitha and Javidi [6] for binary hypothesis testing to binary classification where generating distributions are unknown under each hypothesis. In the same spirit of [6], our two-phase test bridges over the fixed-length test by Gutman [11] and the sequential test by Haghifam et al. [12, Eq. (7)] by having error exponents close to either one with proper choices of parameters $\gamma$. Specifically, under any pair of distributions $(P_1, P_2)$, if the target value $\gamma$ satisfies $\gamma > \min\{D_{\mu\pi}(P_1||P_2), D_{\mu\pi}(P_2||P_1)\} =: \gamma(P_1, P_2, \alpha)$, we have $\lambda_2 = 0$ and our test reduces to Gutman’s test. This is because the maximum exponent of the excess-length probability $P_{\tau|\tau > n}$ that our two-phase test can ensure is $D_{\mu\pi}(P_1||P_2)$ under hypothesis $H_1$. In another extreme when $\gamma \rightarrow 0$, $\lambda_1$ tends to $\text{GJS}(P_1, P_2, \alpha)$. With proper choice of $k$, the achievable exponents of our approach the exponents of the sequential test in [12, Theorem 2]. For $\gamma \in (0, \gamma(P_1, P_2, \alpha))$, our two-phase test has performance in between the fixed-length and the sequential tests.

The above discussions also hold under the Bayesian setting considered in [12, Section II.C], where the authors showed that sequential tests outperformed Gutman’s test significantly in terms of the Bayesian error exponent. Here we follow the same setting as [12, Section II.C] by assigning prior probabilities $\pi_1 \in (0,1)$ and $\pi_2 = 1 - \pi_1$ for $H_1$ and $H_2$ respectively. The Bayesian error probability is

$$P_{\text{Bayesian}}(\Phi) = \pi_1 \beta_1(\Phi|P_1, P_2) + \pi_2 \beta_2(\Phi|P_1, P_2),$$

and the corresponding exponent is defined as

$$E_{\text{Bayesian}}(\Phi|P_1, P_2) = \liminf_{n\rightarrow\infty} \frac{-\log P_{\text{err}}(\Phi)}{n}.$$  

The denominator here is set to $N$ in order to maintain the same parameters as in [12, Eq. (14)] for comparison with the numerical results. For fair comparisons, in both our and Gutman’s test, the test sample size $n$ is set as $n = N/\min\{\beta_1^*, \beta_2^*\}$.

For our two-phase test, under any pair of distributions $(P_1, P_2)$, the achievable Bayesian error exponent satisfies

$$E_{\text{Bayesian}}(\Phi_{tp}|P_1, P_2) \geq \max_{\lambda > 0, (\lambda_1, \lambda_2) \in G(\gamma)} \min \{ \min\{\lambda_1, k\lambda + \gamma\}, \min\{\lambda_2, H(k, \alpha, \lambda | P_1, P_2) + \gamma\} \},$$

where the set $G(\gamma)$ is defined as follows:

$$G(\gamma) = \{(\lambda_1, \lambda_2) : \min_{i\in[2]} F_i(\alpha, \lambda_i|P_1, P_2) \geq \gamma \}.$$  

In Fig. 2, we plot the Bayesian error exponent of our test with different parameters and compare our results with the corresponding Bayesian error exponent of Gutman’s test and the sequential test in [12, Theorem 2]. For our two-phase test, the extreme case of $k = 1$ and appropriately chosen $\gamma^*$ corresponds to a fixed-length test and another extreme case of $\gamma \rightarrow 0$ and $k$ sufficient large corresponds to a sequential test.

Finally, we discuss the influence of $\alpha$, the ratio between the lengths of training sequences and the length of the test sequence on the performance of our two-phase test. Note that both $F_i(\alpha, \lambda_i|P_1, P_2), i \in [2]$ and $H(k, \alpha, \lambda | P_1, P_2)$ increase in $\alpha$ and thus the more the training sequences (larger $\alpha$), the better the performance. We first consider the extreme case of $\alpha \rightarrow \infty$, i.e., when the training sequence is unlimited so that generating distributions $(P_1, P_2)$ are estimated accurately, for each $i \in [2]$, we have

$$F_i(\alpha, \lambda_i|P_1, P_2) = \min_{Q_{i} \in \mathcal{P}(\mathcal{X}) : D(Q_{i}||P_1) \leq \lambda_i} D(Q_{i}||P_1) = D(Q_{i}||P_1),$$

(46)
where $Q^*_i$ satisfies

$$Q^*_i = \frac{P_i(x)^{r_{1,i}^*} P_i(x)^{r_{2,i}^*}}{\sum_{a \in X} P_i(x)^{r_{1,a}^*} P_i(x)^{r_{2,a}^*}},$$

and $(r_{1,i}^*, r_{2,i}^*)$ are Lagrange multipliers such that $D(Q^*_i || P_2) = \lambda_2$ and $D(Q^*_2 || P_1) = \lambda_1$ respectively. Furthermore, we have

$$H(k, \infty, \lambda | P_1, P_2) = \min_{Q \in P(X)} kD(Q || P_2) = kD(Q^* || P_2),$$

where $Q^*$ satisfies

$$Q^* = \frac{P_i(x)^{r^*} P_i(x)^{r^*}}{\sum_{a \in X} P_i(x)^{r^*} P_i(x)^{r^*}},$$

and $r^*$ is the Lagrange multiplier such that $D(Q^* || P_1) = \lambda$. Then we have the type-I and type-II error exponents satisfy

$$E_1(\Phi_{tp} | P_1, P_2) = \min \{ D(Q^*_2 || P_1), kD(Q^* || P_1) + \gamma \},$$

$$E_2(\Phi_{tp} | P_1, P_2) = \min \{ D(Q^*_1 || P_2), kD(Q^* || P_2) + \gamma \},$$

where $(Q^*_1, Q^*_2)$ satisfy $\min_{i \in [2]} D(Q^*_i || P_i) > \gamma$. In the other extreme case of $\alpha \to 0$, we find that $H(k, \alpha, \lambda | P_1, P_2) = 0$ and $E_i(\alpha, \lambda) | P_1, P_2) = 0$ for each $i \in [2]$. The influence of $\alpha$ on the Bayesian error exponent follow in a similar manner.

IV. ALMOST FIXED-LENGTH CLASSIFICATION OF MULTIPLE HYPOTHESES

In this section, we generalize our results for binary classification in the previous section to the general case of $M$-ary classification to account for more than two decision outcomes in practical applications.

A. Problem Formulation

In $M$-ary classification, we are given $M$ training sequences $X^N := (X_1^N, \ldots, X_M^N) \in (X^N)^M$ of length $N \in \mathbb{N}$ generated i.i.d. according to distinct unknown distributions $P = (P_1, \ldots, P_M) \in \mathcal{P}(X)^M$ and a test sequence $Y^\tau$ generated i.i.d. according to one of the $M$ distributions, where $\tau$ is a random stopping time with respect to the filtration $\sigma\{X^N, Y_1, \ldots, Y_\tau\}$. The task is to design a test $\Phi = (\tau, \phi_\tau)$, which consists of a random stopping time $\tau$ and a mapping $\phi_\tau: X^{MN} \times X^\tau \to \{H_1, H_2, \ldots, H_M\}$, to decide among the following $M$ hypotheses:

- $H_j$: the test sequence $Y^\tau$ is generated i.i.d. from the same distribution as the $j^{th}$ training sequence $X_j^N, j \in [M]$.

Given any test $\Phi$ and any tuple of distributions $P \in \mathcal{P}(X)^M$, we have the following $M$ error probabilities to evaluate the performance of the test:

$$\beta_j(\Phi | P) = P \{ \phi(\Phi, X^N, Y^\tau) \neq H_j \}, j \in [M],$$

(51)
where for each \( j \in [M] \), we define \( \mathbb{P}_j \{ \cdot \} := \Pr \{ \cdot | H_j \} \) under which \( X_j^N \sim P_j \) for all \( i \in [M] \) and \( Y^T \sim P_j \). Similar to binary classification, we are interested in deriving the type-\( j \) error exponent for each \( j \in [M] \):

\[
E_j(\Phi|P) := \liminf_{N \to \infty} \frac{-\log \beta_j(\Phi|P)}{E[\tau]},
\]

and Bayesian error exponent:

\[
E_{\text{Bayesian}}(\Phi|P) := \liminf_{n \to \infty} \frac{-\log \sum_{j=1}^M \pi_j \beta_j(\Phi|P)}{E[\tau]},
\]

where \( \pi_j \in (0, 1) \) is the prior probability of hypothesis \( H_j \) for each \( j \in [M] \) such that \( \sum_{j=1}^M \pi_j = 1 \).

Similarly to binary classification, the achievable error exponents of a fixed-length test that achieves the same exponent as Gutman’s test, Gutman’s test is the smallest reject region.

We first recall the results of the fixed-length test by Gutman [11]. For any positive real number \( \lambda \in \mathbb{R}_+ \), Gutman’s test \( \Phi_{\text{Gut}}^{(M)} = (n, \phi_{n, \text{Gut}}^{(M)}) \) has a fixed stopping time \( \tau = n \) and proceeds as follows using the training and test sequences \((X^N, Y^n)\):

\[
\phi_{n, \text{Gut}}^{(M)}(X^N, Y^n) = \begin{cases} 
H_1 & \text{if } \text{GJS}(\hat{T}_X^N, \hat{T}_Y^n, \alpha) \geq \lambda, \forall i \in [2, M], \\
H_j & \text{if } \text{GJS}(\hat{T}_X^N, \hat{T}_Y^n, \alpha) < \lambda, \exists i \in [1, M] \\
& \text{and } \text{GJS}(\hat{T}_X^N, \hat{T}_Y^n, \alpha) \geq \lambda, \forall t \in [M]\setminus\{i\}, \\
H_\ast & \text{otherwise,}
\end{cases}
\]

where \( H_\ast \) denotes the reject option, which indicates that a reliable decision could not be made.

Gutman proved the following result.

**Theorem 7.** For any tuple of distributions \( P \in \mathcal{P}(\mathcal{X})^M \) and any \( \lambda \in \mathbb{R}_+ \), Gutman’s test satisfies

\[
E_j(\Phi_{\text{Gut}}^{(M)}|P) \geq \lambda, j \in [M].
\]

Furthermore, among all fixed-length tests that achieve the same exponent as Gutman’s test, Gutman’s test the smallest reject region.

Note that the largest achievable exponent \( \lambda \in \mathbb{R}_+ \) of Gutman’s test such that the rejection probability vanishes as \( n \to \infty \) is

\[
\hat{\lambda} = \min_{(i,j) \in \mathcal{M}_{\text{dis}}} \text{GJS}(P_i, P_j, \alpha),
\]

where \( \mathcal{M}_{\text{dis}} \) was defined as the set \( \{(i,j) \in [M]^2 : i \neq j\} \).

We then recall the achievable error exponents for the sequential test by Haghifam, Tan and Khisti [12 Theorem 5]. The sequential test \( \Phi_{\text{seq}}^{(M)} = (\tau_{\text{seq}}^{(M)}, \phi_{\tau_{\text{seq}}^{(M)}}^{(M)}) \) consists of the random stopping time

\[
\tau_{\text{seq}} = \min \left( N^2, \inf \left\{ n \in \mathbb{N} : |\Psi_n| \geq M - 1 \right\} \right),
\]

and the decision rule

\[
\phi_{\tau_{\text{seq}}^{(M)}} = H_j, j = [M]\setminus\{\tau_{\text{seq}}\},
\]

where the set \( \Psi_n \) is defined as

\[
\Psi_n \triangleq \left\{ i \in [M] : \exists 1 \leq k \leq n \text{ such that } k \text{GJS}(\hat{T}_X^N, \hat{T}_Y^n, N/k) > \mu N \right\}.
\]

Haghifam, Tan and Khisti proved the following result.

**Theorem 8.** Consider any tuple of distributions \( P \in \mathcal{P}(\mathcal{X})^M \) and any \( \mu \in [0, \min_{(i,j) \in \mathcal{M}} C(P_i, P_j)]^+ \) Let \( \theta_{\max}^{(i),\gamma} \) be the solution of the following equality

\[
\text{GJS}(P_i, P_j, \theta_{\max}^{(i),\gamma}) = \mu \theta_{\max}^{(i),\gamma}, \forall (i,j) \in \mathcal{M}_{\text{dis}}.
\]

Let \( \theta_{\max}^{(i),\gamma} \) be the solution of the following equality

\[
\text{GJS}(P_i, P_j, \theta_{\max}^{(i),\gamma}) = \mu \theta_{\max}^{(i),\gamma}, \forall (i,j) \in \mathcal{M}_{\text{dis}}.
\]

For each \( j \in [M] \), the sequential test satisfies

\[
E_j(\Phi_{\text{seq}}^{(M)}|P) \geq \min_{i \in \mathcal{M}_j} \text{GJS}(P_i, P_j, \theta_{\max}^{(i),\gamma}),
\]

where \( \mathcal{M}_j \) was defined as the set \( \{i \in [M] : i \neq j\} \).

\[\text{The size of } \Psi_{\tau_{\text{seq}}} \text{ is } M - 1, \text{ thus the sequential test is reasonable.}\]

\[\text{The Chernoff information } C(\cdot) \text{ was defined in Eq. (40).}\]
Theorems 7 and 8 imply that the sequential test in Eq. (24) achieves different error exponent under each hypothesis while Gutman’s test achieves the same error exponents under each hypothesis with an additional reject option. Furthermore, Haghifam, Tan and Khisti [12, Theorem 6] demonstrated that the achievable Bayesian error exponent of the sequential scheme is equal to that of Gutman’s test, i.e., $E_{\text{Bayesian}}(\Phi_{\text{Gut}}^{(M)}|\mathbf{P}) = E_{\text{Bayesian}}(\Phi_{\text{seq}}^{(M)}|\mathbf{P})$. Note that the sequential test in Eq. (24) does not need the additional reject option to achieve the same exponent as Gutman’s test with the reject option. However, the sequential test has high computation complexity as discussed for hypothesis testing and binary classification in previous sections.

We wonder whether we could achieve the performance close to the sequential test $\Phi_{\text{seq}}^{(M)}$ with a test that does not need a reject option and has asymptotic complexity of Gutman’s test. In the next two subsections, we propose such a test and prove its desired performance.

C. Our Two-phase Test

We need the following definitions to present our test. Given training sequences $\mathbf{X}^N$ and a test sequences $\mathbf{Y}^n$, define the following function that denote the indices of the minimum scoring values

$$i^*(\mathbf{X}^N, \mathbf{Y}^n) = \arg\min_{i \in [M]} \text{GJS}(\tilde{T}_{X^N}, \tilde{T}_{Y^n}, \alpha),$$

and the set $\mathcal{I} = [M] \setminus i^*(\mathbf{X}^N, \mathbf{Y}^n)$.

Consider any integer $n \in \mathbb{N}$ and any positive real numbers $(k, \lambda_1, \ldots, \lambda_M) \in \mathbb{R}_+^{M+1}$, our test is an almost fixed-length test with two phases. Specifically, the random stopping time $\tau^{(M)}$ of our test satisfies

$$\tau^{(M)} = \begin{cases} n & \text{if } \forall i \in \mathcal{I} \text{ GJS}(\tilde{T}_{X^N}, \tilde{T}_{Y^n}, \alpha) \geq \lambda_i, \\ kn & \text{otherwise}, \end{cases}$$

(62)

The decision rules of our two-phase test are the same equation as the binary case in Eq. (13) and Eq. (14) except that $j$ is replaced by $i^*(\mathbf{X}^N, \mathbf{Y}^n)$ and $i^*(\mathbf{X}^N, \mathbf{Y}^kn)$ at the stopping time $n$ and $kn$, respectively.

For simplicity, we use $\Phi_{\text{tp}}^{(M)}$ to denote the above two-phase test of $M$-ary classification. Note that our setup does not include the rejection option in the final decision.

D. Main Results and Discussions

For simplicity, let $\mathcal{Q} = (Q_1, Q_2, Q_3) \in \mathcal{P}(\mathcal{X})^3$. Given any thresholds $\lambda^M = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}_+^M$, for each $j \in [M]$, define the following exponent function

$$F_j(\alpha, \lambda^M|\mathbf{P}) = \min_{(i,l) \in \mathcal{M}_{\text{dis}}} \min_{\mathcal{Q} \in \mathcal{P}(\mathcal{X})^3} D(\alpha, Q_i, P_i, P_j),$$

(63)

where $D(\alpha, Q_i|P_i, P_j) = \alpha D(Q_1||P_1) + \alpha D(Q_2||P_1) + \alpha D(Q_3||P_j)$ is the linear combination of three KL divergence terms. As we shall show, $F_j(\alpha, \lambda^M|\mathbf{P})$ is critical to bound the probability that our test proceeds the second phase, i.e., $\Pr\{\tau > n\} \leq \exp(-n\gamma)$ for some targeted exponent $\gamma \in \mathbb{R}_+$.

Note that for each $j \in [M]$, $F_j(\alpha, \lambda^M|\mathbf{P})$ is convex and non-increasing in $\lambda_j$. In particular, $F_j(\lambda^M|\mathbf{P}) = 0$ if $\lambda^M = (\lambda_1, \ldots, \lambda_M)$ satisfies that $\lambda_i \geq \text{GJS}(P_i, P_j, \alpha)$ and $\lambda_i \geq \text{GJS}(P_j, P_j, \alpha)$ for some distinct pair $(i,l) \in \mathcal{M}_{\text{dis}}$ (cf. (15)). Furthermore, when $\lambda^M$ is a all zero vector, $F_j(\alpha, \lambda^M|\mathbf{P})$ achieves the following maximum value:

$$\tilde{\gamma}_j(\alpha|\mathbf{P}) := \min_{(i,l) \in \mathcal{M}_{\text{dis}}} \min_{\mathcal{Q} \in \mathcal{P}(\mathcal{X})^3} D(\alpha, Q_i, Q_j||P_i, P_j, P_j).$$

(64)

Finally, for each $j \in [M]$, define another exponent function

$$L_j(k, \alpha|\mathbf{P}) = \min_{i \in [M]} \min_{\mathcal{Q} \in \mathcal{P}(\mathcal{X})^3} U(k,\alpha, Q_i|P_i, P_j),$$

(65)

where $U(k, \alpha, Q_i|P_i, P_j) = \alpha D(Q_1||P_i) + \alpha D(Q_2||P_i) + kD(Q_3||P_j)$ is analogous to $D(\alpha, Q_i|P_i, P_j, P_j)$. As we shall show, $L_j(k, \alpha|\mathbf{P})$ characterizes the achievable type-$j$ error exponent in the second phase of our test.

With these definitions, our result states as follows.

**Theorem 9.** For any $(k, \gamma) \in \mathbb{R}_+^2$, under any tuple of distributions $\mathbf{P} \in \mathcal{P}(\mathcal{X})^M$, the achievable type-$j$ error exponent of our two-phase test satisfies that for each $j \in [M]$,

$$E_j(\Phi_{\text{tp}}^{(M)}|\mathbf{P}) \geq \min \{\lambda_j, L_j(k, \alpha|\mathbf{P}) + \gamma\},$$

(66)
where the thresholds $\lambda^M = (\lambda_1, \ldots, \lambda_M)$ satisfy $\lambda^M \in \tilde{G}(\gamma) := \{\tilde{\lambda}^M \in \mathbb{R}_+^M : \min\limits_{j \in [M]} F_j(\alpha, \tilde{\lambda}^M|P) \geq \gamma\}$. Furthermore, the Bayesian error exponent of our two-phase test satisfies
\[
E_{\text{Bayesian}}(\Phi_{tp}^{(M)}|P) \geq \max_{\lambda^M \in \tilde{G}(\gamma)} \min_{j \in [M]} \left\{ \lambda_j, L_j(k, \alpha|P) + \gamma \right\}.
\]
(67)

The proof of Theorem 9 is provided in Appendix C. We make several remarks.

Firstly, Theorem 9 generalizes the results of almost fixed-length binary classification in the previous section to classification with more than two decision outcomes in both the Neyman-Pearson and Bayesian settings. In Neyman-Pearson setting, our two-phase test bridges over the fixed-length test by Gutman [11] and the sequential test by Haghigham et al. [12, Eq. (24)] by having error exponents close to either one with proper choices of parameters $\gamma$ and $k$. Specifically, under any tuple of distributions $P$, if the target value $\gamma$ satisfies $\gamma > \min\limits_{j \in [M]} \tilde{\gamma}_j(\alpha|P) =: \gamma(P, \alpha)$, the set $\tilde{G}(\gamma)$ contains only the all zero vector, and thus the only valid thresholds $\lambda^M$ of our test are all zero. In this case, our two-phase test reduces to a fixed-length test using $n$ samples (cf. (62)). In another extreme when $\gamma \to 0$, the maximal threshold $\lambda_j$ tends to $\min_{(i,l) \in M_{\text{dis}}} \max\{\text{GJS}(P_i, P_j, \alpha), \text{GJS}(P_i, P_j, \alpha)\}$. With proper choice of $k$, the achievable type-$j$ exponent of our test approach that of the sequential test in [12, Theorem 5]. For $\gamma \in (0, \gamma(P, \alpha))$, our two-phase test has performance in between the fixed-length and the sequential tests.

Recall that the sequential test $\Phi_{seq}^{(M)}$ and the fixed-length test $\Phi_{Gut}^{(M)}$ achieves the same Bayesian error exponent for $M$-ary classification [12, Theorem 6]. However, the fixed-length test $\Phi_{Gut}^{(M)}$ requires an additional reject option. In contrast, with proper choices of design parameters, our two-phase test could achieve roughly the same Bayesian error exponent without a reject option at the asymptotic complexity of the fixed-length test $\Phi_{Gut}^{(M)}$. To illustrate our result, in Fig. 3, we plot the achievable Bayesian error exponents of our two-phase test and the sequential test in [12, Eq. (24)].

Finally, we discuss the influence of $\alpha$, the ratio between the lengths of training sequences and the length of the test sequence on the performance of our two-phase test. Similar to binary classification, for $j \in [M]$, both $F_j(\alpha, \lambda^M|P)$ and $L_j(k, \alpha|P)$ increase in $\alpha$ and thus the more the training sequences (larger $\alpha$), the better the performance. Consider the extreme case of $\alpha \to \infty$. In this case, the training sequence is unlimited so that generating distributions $P$ are estimated accurately. For each $j \in [M]$, we have
\[
F_j(\infty, \lambda^M|P) = \min_{(i,l) \in M_{\text{dis}}} \min_{Q \in \mathcal{P}(\mathcal{X}) : D(Q||P_i) \leq \lambda_i, D(Q||P_j) \leq \lambda_j} D(Q||P_j)
\]
(68)
\[
= \Gamma_j(\lambda^M|P),
\]
(69)
and
\[
L_j(k, \infty|P) = \min_{i \in M_j} \min_{Q \in \mathcal{P}(\mathcal{X}) : D(Q||P_i) < D(Q||P_j)} k D(Q||P_j)
\]
(70)
\[
= \Omega_j(k|P).
\]
(71)

This implies that the performance of our two-phase test for $M$-ary classification approaches that for multiple hypothesis testing when $\alpha \to \infty$. 

![Fig. 3. Illustration of the Bayesian error exponent of $M$-ary classification problem for distributions $P_1 = [0.3, 0.3, 0.4], P_2 = [0.4, 0.5, 0.1]$ and $P_3 = [0.1, 0.7, 0.2]$. Note that the exponent here is calculated from Eq. (43).](image-url)
V. Conclusion

We proposed two-phase tests, analyzed the achievable error exponents and showed that our two-phase tests bridged over fixed-length and sequential tests for $M$-ary hypothesis testing, binary classification and $M$-ary classification. For each case, we showed that with proper design parameters, our test achieved performance close to the sequential test at the asymptotic complexity of a fixed-length test and thus strode a good tradeoff between design complexity and achievable performance. Our results significantly generalized the results of Lalitha and Javidi [9] for binary hypothesis testing by considering multiple decision outcomes and considering the more practical scenarios where the generating distribution under each hypothesis is unknown but available from training sequences. We illustrated our result for each case with numerical examples and analytical discussions. In particular, we discussed the influence of the ratio of the lengths of training sequences and testing sequences on the performance of our tests when the generating distribution is unknown.

There are several avenues for future research. Firstly, we only derived achievability results in this paper without a converse. Thus, one can derive matching converse results under certain criterion such as the generalized Neyman-Pearson criterion [11] usually adopted for hypothesis testing with unknown generating distributions. Secondly, we only considered discrete alphabet sequences. It is of interest to generalize our results to continuous sequences. In this case, the method of types that we used failed and novel ideas such as kernel methods might be required [24]. Finally, we focused on hypothesis testing and statistical classification in this paper. We believe that two-phase tests would also be applicable to other statistical inference problems, e.g., distributed detection problem [18], [25], quickest change-point detection problem [19], [26] and outlier hypothesis testing [20], [21], [27].

APPENDIX

A. Proof of Theorem 3

Given any $P \in \mathcal{P}(X)^M$, the type-$j$ error probability can be calculated as follows: for each $j \in [M]$

$$\beta_j(M) = P_j \{ \tau = n, (\phi_n(Y^n))_j \neq H_j \} + P_j \{ \tau = kn, (\phi_{kn}(Y^{kn}))_j \neq H_j \}. \tag{72}$$

For the first term, it can be upper bound as follows:

$$P_j \{ \tau = n, (\phi_n(Y^n))_j \neq H_j \} \tag{73}$$
$$= P_j \{ \forall i \in N, D(\tilde{T}_{Y^n}P_i) > \lambda_i \text{ and } j \neq \arg \min_{i \in [M]} D(\tilde{T}_{Y^n}P_i) \} \tag{74}$$
$$\leq P_j \{ D(\tilde{T}_{Y^n}P_j) > \lambda_j \} \tag{75}$$
$$\leq \sum_{y^n : D(\tilde{T}_{Y^n}P_j) > \lambda_j} P^n(jy^n) \tag{76}$$
$$\leq \sum_{Q \in \mathcal{P}(X) : D(Q||P_j) > \lambda_j} P^n(Q) \tag{77}$$
$$\leq \sum_{Q \in \mathcal{P}(X) : D(Q||P_j) > \lambda_j} \exp \{-nD(Q||P_j)\} \tag{78}$$
$$\leq \exp \{-n\lambda_j + \log(n+1)\}, \tag{79}$$

where (78) follows from the upper bound of the probability of a type class.

For the second term, i.e., $P_j \{ \tau = kn, (\phi_{kn}(Y^{kn}))_j \neq H_j \}$, it can be decomposed as the excess-length probability $P_j \{ \tau = kn \}$ and the error probability $P_j \{ (\phi_{kn}(Y^{kn}))_j \neq H_j \}$. Recall the definition of $M_{dis}$ in Eq. (15). Under hypothesis $H_j$, the excess-length probability equals to

$$P_j \{ \tau = kn \} \tag{80}$$
$$= P_j \{ \exists i \in N, D(\tilde{T}_{Y^n}P_i) \leq \lambda_i \} \tag{81}$$
$$= P_j \{ \exists (i,l) \in M_{dis}, \text{ s.t. } D(\tilde{T}_{Y^n}P_i) \leq \lambda_i, D(\tilde{T}_{Y^n}P_l) \leq \lambda_l \} \tag{82}$$
$$\leq \sum_{(i,l) \in M_{dis}} \sum_{y^n : D(\tilde{T}_{Y^n}P_i) \leq \lambda_i, D(\tilde{T}_{Y^n}P_l) \leq \lambda_l} P^n(Q) \tag{83}$$
$$= \sum_{(i,l) \in M_{dis}} \sum_{Q \in \mathcal{P}(X) : D(Q||P_i) \leq \lambda_i, D(Q||P_l) \leq \lambda_l} P^n(Q) \tag{84}$$
$$\leq \sum_{(i,l) \in M_{dis}} \sum_{Q \in \mathcal{P}(X) : D(Q||P_i) \leq \lambda_i, D(Q||P_l) \leq \lambda_l} \exp \{-nD(Q||P_j)\} \tag{85}$$
$$\leq M(M-1) \max_{(i,l) \in M_{dis}} \sum_{Q \in \mathcal{P}(X) : D(Q||P_i) \leq \lambda_i, D(Q||P_l) \leq \lambda_l} \exp \{-nD(Q||P_j)\} \tag{86}.$$
\[
M(M - 1)|P^n(\mathcal{X})| \max_{(i,j) \in M} \max_{Q \in P^m(\mathcal{X}) : D(Q||P) \leq \lambda_i, D(Q||P_j) \leq \lambda_i} \exp \left\{ -nD(Q||P_j) \right\} \\
\leq \exp \left\{ -n\Gamma_j(\lambda_i^M[P] + \log M(M - 1) + \lambda_i \log(2n + 1) \right\},
\]
where (88) follows from the definition of \( \Gamma_j(\lambda^M[P]) \) in Eq. (46) and the fact that the number of the set of types of length \( n \) is \( |P^n(\mathcal{X})| = (n + 1)^{|\mathcal{X}|} \). Therefore, for each \( j \in [M] \) and \( \lambda^M = (\lambda_1, \ldots, \lambda_M) \in \hat{G}(\gamma) \), we have
\[
\lim \inf_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_j \{ \tau = k\tau \} \geq \Gamma_j(\lambda_i^M[P]) \geq \gamma,
\]
where \( \hat{G}(\gamma) = \{ \lambda^M \in \mathbb{R}^M : \inf_{i \in [M]} \Gamma_j(\lambda^M[P]) \geq \gamma \} \).

Recall the definition of \( M_j = \{ i \in [M] : i \neq j \} \). Under hypothesis \( H_j \), the error probability of the second phase, i.e., \( \mathbb{P}_j \{ \phi_{kn}(Y^{kn}) \neq H_j \} \), can be upper bound as
\[
\mathbb{P}_j \{ \phi_{kn}(Y^{kn}) \neq H_j \} = \mathbb{P}_j \{ j \neq \arg \min_{i \in [M]} D(\hat{T}_{Y^{kn}}||P_i) \} \\
= \mathbb{P}_j \{ \exists i \in M_j, \text{ s.t. } D(\hat{T}_{Y^{kn}}||P_i) < D(\hat{T}_{Y^{kn}}||P_j) \} \\
\leq \sum_{i \in M_j} \mathbb{P}_j \{ D(\hat{T}_{Y^{kn}}||P_i) < D(\hat{T}_{Y^{kn}}||P_j) \} \\
= \sum_{i \in M_j} \sum_{y^{kn} : D(\hat{T}_{Y^{kn}}||P_i) < D(\hat{T}_{Y^{kn}}||P_j)} P_j(y^{kn}) \\
= \sum_{i \in M_j} \sum_{Q \in P^m(\mathcal{X}) : D(Q||P_i) < D(Q||P_j)} \exp \{ -knD(Q||P_i) \} \\
\leq M \max_{i \in M_j} \sum_{Q \in P^m(\mathcal{X}) : D(Q||P_i) < D(Q||P_j)} \exp \{ -knD(Q||P_j) \} \\
= \exp \{ -n\Omega_j(k[P] + \log M + |\mathcal{X}| \log(2n + 1) \}.
\]

When \( \gamma > 0 \), we have \( \mathbb{E}[\tau] = n + kn \cdot \exp \{ -n\gamma \} \), which satisfies that \( \lim_{n \to \infty} \frac{\mathbb{E}[\tau]}{n} = 1 \). Combining the results in Eq. (79), Eq. (89) and (99), given any thresholds \( \lambda^M = (\lambda_1, \ldots, \lambda_M) \in \hat{G}(\gamma) \), for each \( j \in [M] \), the type-\( j \) exponent satisfies
\[
E_j(\Phi_{ht}^M[P]) \geq \min \{ \lambda_j, \Omega_j(k[P] + \gamma) \}.
\]
where Then the Bayesian error exponent of the two-phase test for multiple hypothesis testing satisfies
\[
E_{\text{Bayesian}}(\Phi_{ht}^M[P]) \geq \max_{\lambda^M \in \hat{G}(\gamma)} \min_{j \in [M]} \{ \lambda_j, \Omega_j(k[P] + \gamma) \}.
\]

**B. Proof of Theorem 6**

Given any \( (P_1, P_2) \in \mathcal{P}(\mathcal{X})^2 \), for each \( i \in [2] \), the error probability under hypothesis \( H_i \) satisfies
\[
\beta_i(\Phi_{ht}^i[P_1, P_2]) = \mathbb{P}_i \{ \tau = n, \phi_k(X_1^n, X_2^n, Y^n) \neq H_i \} + \mathbb{P}_i \{ \tau = kn, \phi_{kn}(X_1^n, X_2^n, Y^{kn}) \neq H_i \}.
\]

For each \( i \in [2] \) and any \( (\lambda_1, \lambda_2) \in \mathbb{R}^2_+ \), define the set \( \mathcal{R}_i = \{ (Q_1, Q_2) \in P^m(\mathcal{X}) \times P^m(\mathcal{X}) : \text{GJS}(Q_1, Q_2, \alpha) \leq \lambda_i \} \) and let \( \mathcal{R}_i^c \) denote the complementary set of \( \mathcal{R}_i \). Under hypothesis \( H_i \), the first term of Eq. (102) is upper bounded as follows:
\[
\mathbb{P}_1 \{ \tau = n, \phi_k(X_1^n, X_2^n, Y^n) \neq H_i \} = \mathbb{P}_1 \{ \text{GJS}(\hat{T}_{X_1^n}, \hat{T}_{Y^n}, \alpha) > \lambda_i \} \\
= \sum_{x^n, y^n : \text{GJS}(\hat{T}_{X_1^n}, \hat{T}_{Y^n}, \alpha) > \lambda_i} P_1(x^n) P_k(y^n) \\
= \sum_{(Q_1, Q_2) \in \mathcal{R}_i^c} P_1^N(T_1^N) P_k^N(T_2^N) \\
\leq \sum_{(Q_1, Q_2) \in \mathcal{R}_i^c} \exp \{ -nD(Q_1||P_1) - nD(Q_2||P_2) \}.
\]
Therefore, for each $i \in [2]$, we have

$$\liminf_{n \to \infty} -\frac{1}{n} \log P_1 \{\tau = n, \phi_n(X_1^n, X_2^n, Y^n) \neq H_i \} \geq \lambda_i. \quad (113)$$

Similarly, under hypothesis $H_2$, we have

$$P_2 \{\tau = n \} \leq \exp \left\{ -n F_2(\alpha, \lambda_2 | P_1, P_2) + |X| \log(N + n + 1) \right\}. \quad (121)$$

Thus, for each $i \in [2]$ and any $(\lambda_1, \lambda_2) \in G(\gamma)$ defined in (45),

$$\liminf_{n \to \infty} -\frac{1}{n} \log P_1 \{\tau = k \} \geq F_i(\alpha, \lambda_i | P_1, P_2) \geq \gamma, \quad (122)$$

where $G(\gamma) = \{(\lambda_1, \lambda_2) : \min_{i \in [2]} F_i(\alpha, \lambda_i | P_1, P_2) \geq \gamma \}$.

Next we upper bound the error probability term, i.e., $P_i \{\phi_kn(X_1^n, X_2^n, Y^{kn}) \neq H_i \}$. Under hypothesis $H_1$, similar to the proof of (103) and (112), we have

$$P_1 \{\phi_kn(X_1^n, X_2^n, Y^{kn}) \neq H_i \} \leq \exp \left\{ -k n \lambda + |X| \log(N + n + 1) \right\}, \quad (123)$$

which demonstrates that the type-I probability in the second phase essentially decay exponentially fast with the speed of the threshold $k \lambda$ when $n$ is sufficiently large.

For any $\lambda \in \mathbb{R}_+$, we define $\mathcal{R} = \{(Q_1, Q_2) \in \mathcal{P}_N(X) \times \mathcal{P}_N(Y) : \text{GJS}(Q_1, Q_2, \alpha) \leq \lambda \}$. Under hypothesis $H_2$, we have

$$\mathbb{P}_2 \{\phi_kn(X_1^n, X_2^n, Y^{kn}) \neq H_2 \}$$

$$= \mathbb{P}_2 \left\{ \text{GJS} \left( \hat{T}_{X^n}, \hat{T}_{Y^{kn}}, \frac{Q}{\lambda} \right) \leq \lambda \right\}$$

$$= \sum_{x^n, y^{kn} : \text{GJS} \left( \hat{T}_{x^n}, \hat{T}_{y^{kn}}, \frac{Q}{\lambda} \right) \leq \lambda} P_1(x_1^n) P_2(y^{kn}) \quad (125)$$
Then the Bayesian error exponent of the two-phase test for binary classification satisfies

$$P_{\lambda} F$$

Given any $$P_{\lambda}$$, the Bayesian error exponent of the two-phase test for binary classification satisfies

$$E_{\text{Bayesian}}(P_{\lambda}|P_1, P_2) \geq \max_{\lambda > 0} \max \left\{ \min \{\lambda_1, k\lambda + \gamma\}, \min \{\lambda_2, H(k, \alpha, \lambda|P_1, P_2) + \gamma\} \right\}.$$ (132)

C. Proof of Theorem

Given any $$P \in \mathcal{P}(\mathcal{X})^M$$, for each $$j \in [M]$$, the achievable type-$$j$$ error probability satisfies

$$\beta_j(P_{\lambda}|P) = P_j\left\{ \tau = n, \phi_{\lambda}(X, Y^n) \neq H_j \right\} + P_j\left\{ \tau = kn, \phi_{\lambda}(X, Y^{kn}) \neq H_j \right\}. $$ (133)

The first term of (133), i.e., the type-$$j$$ error probability of the first phase, can be upper bounded as follows:

$$P_j\left\{ \tau = n, \phi_{\lambda}(X, Y^n) \neq H_j \right\} \leq \sum \left\{ \text{GJS} \left( \hat{T}_{X^n}, \hat{T}_{Y^n}, \alpha \right) \geq \lambda, \forall \hat{k} \neq i^*(X^n, Y^n) \right\} \leq \sum \text{exp} \left\{ -n\lambda_j + |\lambda| \log(N + n + 1) \right\}. $$ (134)

The second term of (133), i.e., the type-$$j$$ error probability of the second phase, can be decomposed as the excess-length probability $$P_j\{\tau = kn\}$$ and the error probability $$P_j\{\phi_{\lambda}(X^n, Y^{kn}) \neq H_j\}$$. Given $$(i, l) \in \mathcal{M}_{\text{dis}}$$ and $$\lambda_1, \ldots, \lambda_M \in \mathbb{R}^+_M$$, define a set $$\mathcal{B} = \{Q \in \mathcal{P}^N(\mathcal{X}) \times \mathcal{P}^N(\mathcal{X}) \times \mathcal{P}^N(\mathcal{X}) : \text{GJS}(Q_1, Q_2, \alpha) \leq \lambda_i, \text{GJS}(Q_2, Q_3, \alpha) \leq \lambda_i \}$$.

Under hypothesis $$H_j$$, the excess-length probability, i.e., $$P_j\{\tau = kn\}$$, can be upper bounded as follows:

$$P_j\{\tau = kn\} \leq \sum \text{exp} \left\{ -n \cdot D(\alpha, Q|P_i, P_l, P_j) \right\} \leq M(M - 1) \max (i, l) \sum \text{exp} \left\{ -n \cdot D(\alpha, Q|P_i, P_l, P_j) \right\} \leq M(M - 1) \max (i, l) \sum \text{exp} \left\{ -n \cdot D(\alpha, Q|P_i, P_l, P_j) \right\}$$ (140)

where (144) follows from the definition of $$F_j(\alpha, \lambda^M|P)$$ in Eq. (63). Thus, for each $$j \in [M]$$ and $$\lambda^M = (\lambda_1, \ldots, \lambda_M) \in \hat{G}(\gamma)$$, we have

$$\liminf_{n \to \infty} \frac{1}{n} \log P_j\{\tau = kn\} \geq F_j(\alpha, \lambda^M|P) \geq \gamma.$$ (146)
where $\tilde{G}(\gamma) = \{\lambda^M \in \mathbb{R}^M : \min_{j \in [M]} F_j(\alpha, \lambda^M | \mathbf{P}) \geq \gamma\}$. Define $\mathcal{S} = \{Q \in \mathcal{P}^N(X) \times \mathcal{P}^N(Y) : \text{GJS}(Q_1, Q_3, \alpha) < \text{GJS}(Q_2, Q_3, \alpha)\}$. Under hypothesis $H_j$, the error probability, i.e., $P_j(\phi_{kn}(X_1^N, Y_1^N, Y^{kn}) \neq H_j)$, is upper bounded as follows:

$$P_j(\phi_{kn}(X_1^N, Y^{kn}) \neq j) = P_j\left\{\exists i \in M_j, \text{s.t.} \text{GJS}(\hat{T}_{X_i^N}, \hat{T}_{Y_i^N, kn}) < \text{GJS}(\hat{T}_{X_i^N}, \hat{T}_{Y_i^N, kn})\right\} \leq \sum_{i \in M_j} P_j\left\{\text{GJS}(\hat{T}_{X_i^N}, \hat{T}_{Y_i^N, kn}) < \text{GJS}(\hat{T}_{X_i^N}, \hat{T}_{Y_i^N, kn})\right\}$$

$$= \sum_{i \in M_j} \sum_{Q \in \mathcal{S}} P_i^n(\hat{T}_{Q_i}^N) P_j^n(\hat{T}_{Q_j}^N) P_j(\gamma^{kn}) \leq \sum_{i \in M_j} \sum_{Q \in \mathcal{S}} \exp \{-nU(\lambda, k, Q|P_i, P_j)\} \leq \exp \{-nL_j(\lambda, k|P) + |\lambda'| \log(N + 1)^2(n + 1) + \log M\}.$$  \hfill (153)

Combining the results in Eq. (136), Eq. (146), and Eq. (153), for any $\lambda^M \in \tilde{G}(\gamma)$, we have

$$E_j(\Phi^{(M)}|P) \geq \min \{\lambda_j, L_j(\lambda, k|P) + \gamma\},$$ \hfill (154)

Then the Bayesian error exponent of the two-phase test for multiple hypothesis testing satisfies

$$E_{\text{Bayesian}}(\Phi^{(M)}|P) \geq \max_{\lambda^M \in \tilde{G}(\gamma)} \min_{j \in [M]} \{\lambda_j, L_j(\lambda, k|P) + \gamma\}. \hfill (155)$$

**REFERENCES**

[1] I. Csiszár and P. Shields, “Information theory and statistics: A tutorial,” *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 417–528, 2004.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[3] H. Chernoff, “A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations,” *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, 1952.

[4] R. Blahut, “Hypothesis testing and information theory,” *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, 1974.

[5] A. Wald and J. Wolfowitz, “Optimum character of the sequential probability ratio test,” *Ann. Math. Statist.*, pp. 326–339, 1948.

[6] A. Lalitha and T. Javidi, “Reliability of sequential hypothesis testing can be achieved by an almost-fixed-length test,” in *IEEE ISIT*, 2016, pp. 1710–1714.

[7] C. C. Leang and D. H. Johnson, “On the asymptotics of M-hypothesis Bayesian detection,” *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 280–282, 1997.

[8] E. Tuncel, “On error exponents in hypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2945–2950, 2005.

[9] C. W. Baum and V. V. Veeravalli, “A sequential procedure for multihypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 40, no. 6, 1994.

[10] V. Dragnia, A. G. Tartakovsky, and V. V. Veeravalli, “Multihypothesis sequential probability ratio tests: I. Asymptotic optimality,” *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2448–2461, 1999.

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

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

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

[14] H.-W. Hsu and I.-H. Wang, “On binary statistical classification from mismatched empirically observed statistics,” in *IEEE ISIT*, 2020, pp. 2533–2538.

[15] N. Merhav and J. Ziv, “A Bayesian approach for classification of Markovs streams,” *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1067–1071, 1991.

[16] S. Saito and T. Matsushima, “Evaluation of error probability of classification based on the analysis of the Bayes code,” in *IEEE ISIT*, 2020, pp. 2510–2514.

[17] J. Unnikrishnan, “Asymptotically optimal matching of multiple sequences to source distributions and training sequences,” *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 452–468, 2015.

[18] H. He, L. Zhou, and Y. F. Tan, “Distributed detection with empirically observed statistics,” *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4349–4367, 2020.

[19] H. He, Q. Zhang, and Y. F. Tan, “Optimal change-point detection with training sequences in the large and moderate deviations regimes,” *IEEE Trans. Inf. Theory*, 2021.

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

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

[22] C. Hsu, C. Li, and I. Wang, “On universal sequential classification from sequentially observed empirical statistics,” Unpublished preprint, http://homepage.ntu.edu.tw/~ihwang/Eprint/ito22usc.pdf, 2022.

[23] W. Hoefding, “Asymptotically optimal tests for multinomial distributions,” *Ann. Math. Statist.*, pp. 369–401, 1965.

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

[25] R. R. Tenney and N. R. Sandell, “Detection with distributed sensors,” *IEEE Trans. Aerosp. Electron. Syst.*, no. 4, pp. 501–510, 1981.

[26] A. N. Pettit, “A non-parametric approach to the change-point problem,” *J. Roy. Stat. Soc. C Appl. Statist.*, vol. 28, no. 2, pp. 126–135, 1979.

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