Analysis on Optimal Error Exponents of Binary Classification for Source with Multiple Subclasses

Hirotó Kuramata † and Hideki Yagi *,†

Department of Computer and Network Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu 182-8585, Tokyo, Japan; h.kuramata@mail.uec.jp
* Correspondence: h.yagi@uec.ac.jp; Tel.: +81-42-443-5366
† These authors contributed equally to this work.

Abstract: We consider a binary classification problem for a test sequence to determine from which source the sequence is generated. The system classifies the test sequence based on empirically observed (training) sequences obtained from unknown sources \( P_1 \) and \( P_2 \). We analyze the asymptotic fundamental limits of statistical classification for sources with multiple subclasses. We investigate the first- and second-order maximum error exponents under the constraint that the type-I error probability for all pairs of distributions decays exponentially fast and the type-II error probability is upper bounded by a small constant. In this paper, we first give a classifier which achieves the asymptotically maximum error exponent in the class of deterministic classifiers for sources with multiple subclasses, and then provide a characterization of the first-order error exponent. We next provide a characterization of the second-order error exponent in the case where only \( P_2 \) has multiple subclasses but \( P_1 \) does not. We generalize our results to classification in the case that \( P_1 \) and \( P_2 \) are a stationary and memoryless source and a mixed memoryless source with general mixture, respectively.

Keywords: binary classification; error exponent; multiple subclasses

1. Introduction

1.1. Background

The problem of learning sources from training sequences and estimating the source from which a test sequence is generated is known as a classification problem. Recently, this problem has been actively studied in a field such as machine learning, and it is desirable to conduct studies that guarantee the performance of the system. In the field of information theory, studies have been conducted mainly to analyze the performance of optimal tests. When the number of sources is two, the binary classification problem can be regarded as a binary hypothesis testing problem, using training sequences. In the setting of binary hypothesis testing, it is assumed that the sources are known, but in real-world applications, the sources are not known in general. Therefore, it is of importance to consider the binary classification problem.

Hypothesis testing includes approaches such as the Bayesian test [1,2] and the Neyman–Pearson test [3–5]. In this paper, we take the latter approach to formulate the best asymptotic error exponent (the exponential part of an error probability).

There are a lot of studies related to the classification problem. We state important points, which are deeply connected to this study in some previous studies. Gutman [3] has shown that type-based (empirical distribution-based) tests asymptotically achieve the maximum type-II error exponent for stationary Markov sources, while the type-I error probability exponentially converges to zero as the length of a test sequence goes to infinity. Zhou et al. [5] derived second-order approximations of the maximum type-I error exponent for stationary and memoryless sources when the type-II error probability is upper bounded by a small constant. On the other hand, for the hypothesis testing problem,
Han and Nomura [6] characterized a first-order maximum error exponent when each sequence is generated from a mixed memoryless source, which is a mixture of stationary and memoryless sources. In addition, they also characterized a second-order maximum error exponent in the case where one source is a stationary and memoryless source and the other source is a mixed memoryless source.

1.2. Contributions

In this paper, we investigate the binary classification problem for stationary memoryless sources with multiple subclasses. The class of sources with multiple subclasses is important in binary classification because there are many such settings in real-world applications. For example, newspaper articles with science headlines consist of topics of physics, chemistry, biology, etc. We assume that sources (subclasses) are characterized by a mixture with some unknown prior distribution (cf. Equation (2)), and the overall sources can be regarded as mixed memoryless sources [6]. The purpose of this paper is to characterize the first- and second-order maximum error exponents in a single-letter form (the term “single-letter form” means an expression which does not depend on lengths of sequences $n$ or $N$ (cf. the formulas for error exponents in Theorems 2–4)).

To this end, we generalize Gutman’s classifier [3], which was shown to be first- and second-order optimal for memoryless sources (with no multiple subclasses) in [5]. This classifier uses training sequences from one of the two sources, as in [3–5], making a type-based decision for a source (subclass) with the smallest skewed Jensen–Shannon divergence [7] among the subclasses. We show that this classifier asymptotically achieves the maximum type-II error exponent in the class of deterministic classifiers for a given pair of distributions when the type-I error probability decays exponentially fast for all pairs of distributions in Theorem 1. We also demonstrate that the structure of this classifier leads to a reversed and more relaxed relation; the maximization of the type-I error exponent when the type-II error probability is upper bounded by a small constant $\epsilon$ ($0 \leq \epsilon < 1$) for sources with multiple subclasses in Theorem 2. In addition, using the Berry–Esseen theorem [8], we derive the second-order maximum error exponent in the case where only one of sources has subclasses in Theorem 3. Finally, the fact that the classifier uses the test sequence from one of two sources motivates us to consider a more general case; the first source is a source with no multiple subclasses, but the second source is given by a general mixture [6,9]. That is, the number of subclasses is not necessarily finite, and the prior distribution of subclasses may not be discrete (cf. Equation (75)). We give characterizations of the first- and second-order maximum error exponents in Theorem 4.

1.3. Related Work

Ziv [10] proposed a classifier based on empirical entropy and discussed the relationship between binary classification and universal source coding. Hsu and Wang [4] characterized the maximum error exponent with mismatched empirically observed statistics. In their achievability proof, a generalization of Gutman’s classifier is also used. Kelly et al. [11] investigated binary classification with large alphabets. Unnikrishnan and Huang [12] investigated the type-I error probability of binary classification using the analysis of weak convergence. Generalizing the binary classification problem, He et al. [13] discussed the binary distribution detection problem, in which a different generalization of Gutman’s classifier is also discussed.

There are also studies which take a Bayesian approach. Merhav and Ziv [1] analyzed the weighted sum of type-I and type-II error probabilities, and subsequently, Saito and Matsushima [2,14] gave a different result via the analysis for the Bayes code.

1.4. Organization

The rest of this paper is organized as follows: In Section 2, we define the notation used in this paper and describe the details of the source and system model. Moreover, we state the problem setting, defining the first- and second-order maximum error exponents.
In Section 3, we first give a classifier which achieves the asymptotically maximum error exponent in the class of deterministic classifiers for sources with multiple subclasses. Next, we characterize the first- and second-order maximum error exponents and the detailed proofs for the first-order representation. In Section 4, we generalize the obtained results to the classification of a mixed memoryless source with general mixture. In Section 5, we present numerical examples. Finally, in Section 6, we provide some concluding remarks and future work.

2. Problem Formulation

2.1. Notation

The set of non-negative real numbers is denoted by $\mathbb{R}^+$. Calligraphic $\mathcal{X}$ stands for a finite alphabet. Upper-case $X$ denotes a random variable taking values in $\mathcal{X}$, and lower-case $x \in X$ denotes its realization. Throughout this paper, logarithms are of base $e$. For integers $a$ and $b$ such that $a \leq b$, $[a, b]$ denotes the set $\{a, a+1, \ldots, b\}$. The set of all probability distributions on a finite set $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. Notation regarding the method of types [15] is as follows: Given a vector $x^n_1 = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$, the type is denoted as $q_{x^n_1}(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i = a\}$, $a \in \mathcal{X}$.

The set of types formed from length-$n$ sequences with alphabet $\mathcal{X}$ is denoted as $\mathcal{P}_n(\mathcal{X})$. The probability that $n$ independent drawings from a probability distribution $Q \in \mathcal{P}(\mathcal{X})$ give $x \in \mathcal{X}^n$ is denoted by $Q(x)$.

2.2. Source with Multiple Subclasses

Consider a source consisting of multiple subclasses. Each subclass is distributed according to a given probability (weight). Let $\{P_i^1\}_{i \in S}$ be a family of probability distributions on a finite alphabet $\mathcal{X}$, where $S = \{1, \cdots, s\}$ is a probability space with probability measure $v(i), i \in S$. That is, the probability of $x \in \mathcal{X}^n$ is given by

$$P_1(x) = \sum_{i=1}^{s} v(i) P_i^1(x),$$

where the $i$-th subclass $P_i^1$ is a stationary and memoryless source. That is, for $x = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$,

$$P_i^1(x) = \prod_{j=1}^{n} P_i^1(x_j)$$

(for notational simplicity, we denote both the multi-letter and single-letter probabilities by $P_i^1$ with a slight abuse of notation).

In view of (2), the sequence $x$ can be regarded as an output from a mixed memoryless source $P_1(\cdot)$, and it is called a test sequence. Similarly, let $\{P_i^2\}_{i \in U}$ be a family of probability distributions, where $U = \{1, \cdots, u\}$ is a probability space with probability measure $w(i), i \in U$. For these mixed sources, if the sources are known, that is, the addressed problem is hypothesis testing, the first- and second-order error exponents were analyzed by Han and Nomura [6]. In this paper, we assume that the sources are unknown and training sequences are available to learn about the source. Sets of training sequences are denoted by $t_1 = \{t_1^1, \cdots, t_1^s\}$ and $t_2 = \{t_2^1, \cdots, t_2^u\}$, where $t_i^j \in \mathcal{X}^N$ of length $N$ is output from
subclass $j$ and $N = \lceil n \gamma \rceil$ for some fixed $\gamma \in \mathbb{R}^+$. Then, the joint probabilities of training sequences are, respectively,

$$P_1(t_1) = \prod_{i=1}^{s} P_i^1(t_{1i}),$$

$$P_2(t_2) = \prod_{j=1}^{u} P_j^2(t_{2j}).$$

We define the class of sources with multiple subclasses on a probability space $S$ with probability measure $v(\cdot)$ as

$$P_S(X) := \{ P = \{ P_i, v(i) \}_{i \in S} : P_i \in \mathcal{P}(X) \},$$

which means $P_1 \in P_S(X)$, where the set of weights $\{ v(i) \}_{i \in S}$ is implicitly fixed. Similarly, we define the class of sources with multiple subclasses on a probability space $U$ with probability measure $w(\cdot)$ as

$$P_U(X) := \{ P = \{ P_i, w(i) \}_{i \in U} : P_i \in \mathcal{P}(X) \},$$

which means $P_2 \in P_U(X)$.

2.3. System Model

The binary classification problem assumed in this paper is shown in Figure 1. It consists of two phases: (I) learning phase and (II) classification phase. We explain the details of each phase.

![Figure 1. System model.](image-url)

(I) Learning phase: Determine the classifier by learning with the training sequences $t_1 = \{ t_{11}, \cdots, t_{1s} \}$ and $t_2 = \{ t_{21}, \cdots, t_{2u} \}$ generated from unknown source $P_1 \in P_S(X)$ and $P_2 \in P_U(X)$, respectively.

(II) Classification phase: It represents the phase in which we judge whether the test sequence $x \in X^n$ generated from $P_1 \in P_S(X)$ or $P_2 \in P_U(X)$ according to the classifier determined in (I).
2.4. Maximum Error Exponent

In this section, we define two error probabilities that arise in a binary classification problem and formulate the maximum error exponents. In the binary classification problem, a test is described as a partition of the space \( \mathcal{X} \times \mathcal{X} \rightarrow \{1, 2\} \). The type-I and type-II error probabilities of a given test \( \phi^n \) are denoted as \( \beta_1(\phi^n|P_1, P_2) \) and \( \beta_2(\phi^n|P_1, P_2) \), respectively. That is,

\[
\beta_1(\phi^n|P_1, P_2) := \mathbb{P}_1\{\phi^n(x, t_1, t_2) = 2\}, \quad \beta_2(\phi^n|P_1, P_2) := \mathbb{P}_2\{\phi^n(x, t_1, t_2) = 1\}.
\]

Here, \( \mathbb{P}_\theta \{\cdot\} \) is the joint probability of training and testing sequences when the underlying parameter is \( \theta \), given by

\[
\mathbb{P}_1\{\phi^n(x, t_1, t_2) = \ell\} := \sum_{(x, t_1, t_2) : \phi^n(x, t_1, t_2) = \ell} P_1(x) \prod_{i=1}^{s} P_1^i(t'_1) \prod_{j=1}^{u} P_1^j(t'_2) \quad \ell \in \{1, 2\},
\]

\[
\mathbb{P}_2\{\phi^n(x, t_1, t_2) = \ell\} := \sum_{(x, t_1, t_2) : \phi^n(x, t_1, t_2) = \ell} P_2(x) \prod_{i=1}^{s} P_2^i(t'_1) \prod_{j=1}^{u} P_2^j(t'_2) \quad \ell \in \{1, 2\}.
\]

We consider the problem of maximizing the type-I error exponent when the type-II error probability is upper bounded by a small constant \( \epsilon \in [0, 1) \). In this study, we characterize the following quantities (the first- and second-order maximum type-I error exponent).

**Definition 1** (First-order maximum error exponent). For any pair of distributions \( P = (P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \) and \( \epsilon \in [0, 1) \), we define

\[
\hat{\lambda}(\epsilon) := \sup_{\lambda \in \mathbb{R}^+} \left\{ \lambda \mid \exists \{\phi^n\}_{n=1}^\infty \text{ s.t. for all sufficiently large } n \right. \\
\beta_1(\phi^n|\hat{P}) \leq \exp(-n\lambda) \quad (\forall \hat{P} \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})), \left. \lim_{n \to \infty} \beta_2(\phi^n|\hat{P}) \leq \epsilon \right\},
\]

where the weights of \( \hat{P} = (\hat{P}_1, \hat{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \) are the same as the weights of \( P = (P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \).

**Definition 2** (Second-order maximum error exponent). For any pair of distributions \( P = (P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \) and \( \epsilon \in [0, 1) \), we define

\[
\hat{\nu}(\epsilon, \lambda) := \sup_{\nu \in \mathbb{R}^+} \left\{ \nu \mid \exists \{\phi^n\}_{n=1}^\infty \text{ s.t. for all sufficiently large } n \right. \\
\beta_1(\nu^n|\hat{P}) \leq \exp(-n\lambda - \sqrt{n\nu}) \quad (\forall \hat{P} \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})), \left. \lim_{n \to \infty} \beta_2(\nu^n|\hat{P}) \leq \epsilon \right\},
\]

where the weights of \( \hat{P} = (\hat{P}_1, \hat{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \) are the same as the weights of \( P = (P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \).

**Remark 1.** In (12) and (13), the type-I error probability is constrained for any \( \hat{P} \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \) for technical reasons. In more detail, this condition is required in the proof of the converse part. This condition was also imposed by Gutman [3], Hsu and Wang [4], and Zhou et al. [5].

In Definitions 1 and 2, we focus on universal tests that perform well for all pairs of distributions with respect to the type-I error probability, and at the same time, constrain the type-II error probability with respect to a particular pair of distributions. We obtain the same result when the weights of \( (\hat{P}_1, \hat{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X}) \) are not fixed.
3. Main Result
3.1. A Test to Achieve Maximum Error Exponent

The rule for estimating whether a test sequence generated from \( P_1 \in \mathcal{P}_S(X) \) or \( P_2 \in \mathcal{P}_U(X) \) is called a decision rule. One of the goals of the classification problem is to design an optimal decision rule which achieves a maximum error exponent based on training sequences. In this section, we present a decision rule that asymptotically achieves the maximum type-II error exponent for any pair of distributions when the type-I error exponent is lower bounded by a constant for all pairs of distributions (cf. Theorem 1).

To define a test that is asymptotically optimum, we define two generalizations of the Jensen–Shannon divergence. These generalizations are related to some variational definitions in [7,16]. For any pair of distributions \((Q_1, Q_2) \in \mathcal{P}(X)^2\) and any number \(\gamma \in \mathbb{R}^+\), let the generalized Jensen–Shannon divergence be

\[
\text{GJS}(Q_1, Q_2, \gamma) := \gamma D\left(Q_1 \left\| \frac{\gamma Q_1 + Q_2}{1 + \gamma} \right\| + D\left(Q_2 \left\| \frac{\gamma Q_1 + Q_2}{1 + \gamma} \right\| \right),
\]

where \(D(p||q)\) denotes the Kullback–Leibler divergence for \(p \in \mathcal{P}(X)\) and \(q \in \mathcal{P}(X)\) defined as

\[
D(p||q) := \sum_{i \in \mathcal{X}} p(i) \log \frac{p(i)}{q(i)}.
\]

The generalized Jensen–Shannon divergence \(\text{GJS}(Q_1, Q_2, \gamma)\) corresponds to a skewed \(\alpha\)-Jensen–Shannon divergence for \(\alpha = \frac{1}{1 + \gamma}\). Additionally, for \((Q_1, Q_2) \in \mathcal{P}_S(X) \times \mathcal{P}(X)\), we define the minimized generalized Jensen–Shannon divergence by

\[
\text{MGJS}(Q_1, Q_2, \gamma) := \min_{\Lambda_i \in S} \text{GJS}(Q_1, Q_2, \gamma).
\]

Given a threshold \(\lambda \in \mathbb{R}^+\) (including \(\lambda = 0\)), the decision rule to achieve the maximum error exponent is given by

\[
\Lambda^\gamma = \left\{ (x, t_1, t_2) \in X^n \times X^{sN} \times X^{uN} \mid \text{MGJS}(q_{t_1}, q_{x}, \gamma) \geq \tilde{\lambda} \right\},
\]

where \(\tilde{\lambda} = \lambda + \eta(n), \eta(n) := \frac{2 \log n + \log |X| \log n (n + N + 1)}{n}\) and \(\Lambda_i^\gamma, i \in \{1, 2\}\) is the set of \((x, t_1, t_2)\) determined to be class \(i\) by the test \(\Lambda_i^\gamma\). By definition, the discriminant function \(\text{MGJS}(q_{t_1}, q_{x}, \gamma)\), appearing on the right-hand side of (17), can also be expressed as

\[
\text{MGJS}(q_{t_1}, q_{x}, \gamma) = \min_{i \in S} \{ \tilde{\gamma} D(q_{t_1} \| q_{y_1}) + D(q_{x} \| q_{y_1}) \},
\]

where \(y_1 := xt_1^i\). From (17) and (18), \(\Lambda_i^\gamma\) is a type-based test and implicitly depends on \(\lambda\). In addition, this test uses training sequences asymmetrically; only sequence \(t_1\) is used, but not \(t_2\) (cf. refs. [3–5]).

**Theorem 1.** For any given \(\lambda \in \mathbb{R}^+\) and any sequence of tests \(\{\phi^n\}\) such that \(\beta_1(\phi^n|\tilde{P}) \leq \exp(-n\lambda), \forall \tilde{P} = (\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(X) \times \mathcal{P}_U(X)\), the sequence of tests \(\{\Lambda_i^\gamma\}\) given by (17) satisfies for any pair of distributions \(P = (P_1, P_2) \in \mathcal{P}_S(X) \times \mathcal{P}_U(X)\),

\[
\beta_1(\Lambda^\gamma|\tilde{P}) \leq \exp(-n\lambda), \forall \tilde{P} \in \mathcal{P}_S(X) \times \mathcal{P}_U(X),
\]

\[
\lim_{n \to \infty} -\frac{1}{n} \log \beta_2(\phi^n|P) \leq \lim_{n \to \infty} -\frac{1}{n} \log \beta_2(\Lambda^\gamma|P).
\]

**Proof.** Equation (19) is derived in Section 3.3.1. The proof of (20) follows from Corollary 1. Although there is a deviation between the exponents in Corollary 1 and for the test \(\Lambda^\gamma\), the deviation vanishes asymptotically.}


Theorem 1 shows that the test $\Lambda^n$ can asymptotically achieve the maximum type-II error exponent among the tests $\phi^n$ for which the type-I error exponent is greater than or equal to $\lambda$. This test also has a reversed and more relaxed property; it achieves $\hat{\lambda}(\epsilon)$, the maximum type-I error exponent when the type-II error probability is upper bounded by a constant $\epsilon$ ($0 \leq \epsilon < 1$) (see the achievability proof of Theorem 2 in Section 3.3.1).

3.2. First-Order Maximum Error Exponent

In this section, we characterize the first-order maximum error exponent in a single-letter form for sources with multiple subclasses.

**Theorem 2.** For any pair of distributions $(P_1, P_2) \in \mathcal{P}_S(X) \times \mathcal{P}_U(X)$, we have

$$\hat{\lambda}(\epsilon) = \sup \left\{ \lambda \left| \sum_{\{j \in U : \text{MGJS}(P_1, P_j^2, \gamma) < \bar{\lambda}\}} w(j) \leq \epsilon \right. \right\}.$$  

(21)

It should be noted that $\hat{\lambda}(\epsilon)$ depends on $\{w(j)\}_{j \in U}$, but not on $\{v(i)\}_{i \in S}$.

**Proof.** The proof is provided in Section 3.3.

**Remark 2.** If $S$ and $U$ are singletons (that is, $s = 1$ and $u = 1$), Theorem 2 reduces to the following formula given by Zhou et al. [5]:

$$\hat{\lambda}(\epsilon) = \text{GJS}(P_1, P_2, \gamma) \quad (0 \leq \epsilon < 1),$$

(22)

which means that $\hat{\lambda}(\epsilon)$ does not depend on $\epsilon$ and the strong converse holds in this case, unlike in the case $|S|, |U| > 1$. On the other hand, for general $S$ and $U$ but in the spacial case of $\epsilon = 0$, formula (21) reduces to

$$\hat{\lambda}(0) = \min_{i \in S, j \in U} \text{GJS}(P_i^1, P_j^2, \gamma).$$

(23)

3.3. Proof of Theorem 2

We divide the proof of Theorem 2 into two parts: the achievability (direct) part and the converse part.

3.3.1. Achievability Part

In the achievability proof, we use the type-based test $\Lambda^n$ given by (17). Fix any

$$\lambda < \sup \left\{ \lambda \left| \sum_{\{j \in U : \text{MGJS}(P_1, P_j^2, \gamma) < \bar{\lambda}\}} w(j) \leq \epsilon \right. \right\}.$$  

(24)

Then, for any pair of distributions $(P_1, P_2) \in \mathcal{P}_S(X) \times \mathcal{P}_U(X)$ and for all pairs of distributions $(\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(X) \times \mathcal{P}_U(X)$, we show

$$\beta_1(\Lambda^n | \tilde{P}_1, \tilde{P}_2) \leq \exp(-n\lambda),$$

(25)

$$\lim_{n \to \infty} \beta_2(\Lambda^n | P_1, P_2) \leq \epsilon.$$  

(26)

First, we prove (25). For preliminaries, we define the following sets used in the proof:

$$\mathcal{T}_n(q_x) := \{x' \in X^n : q_{x'} = q_x\},$$  

(27)

$$\Lambda^n_q := \{(x, t_1) : \text{MGJS}(q_{t_1}, q_x, \gamma) \geq \tilde{\lambda}\},$$  

(28)

$$\Gamma(\Lambda) := \{(x,t_1) : (x, t_1) \in \Lambda\}.$$  

(29)
where $\tilde{\Lambda}_2^n$ is the projection of $A^n \subseteq X^n \times X^{nN} \times X^{nN}$ onto the space $X^n \times X^{nN}$. To evaluate the probability of a source sequence being in $T_n(q_x)$, the following relationship holds from the method of types [15].

**Lemma 1.** Suppose that the sequence $x$ is sampled independently from the source $P \in \mathcal{P}(X)$. Then,

$$
1 \left| \mathcal{P}_n(X) \right| \exp \{-nD(q_x||P)\} \leq P(T_n(q_x)) \leq \exp \{-nD(q_x||P)\}, \quad (30)
$$

where $\mathcal{P}_n(X)$ denotes the set of types formed from length-$n$ sequences with alphabet $X$ and $|\mathcal{P}_n(X)| \leq (n + 1)^{|X|}$.

Then, an upper bound on the type-II error probability of the test $\Lambda^n = (\Lambda_1^n, \Lambda_2^n)$ for all pairs of distributions $(\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(X) \times \mathcal{P}_H(X)$ can be evaluated as follows:

$$
\beta_1(\Lambda^n | \tilde{P}_1, \tilde{P}_2)
= \sum_{(x, t_1, t_2) \in \Lambda^n_2} \tilde{P}_1(x) \prod_{r=1}^{s} \tilde{P}_1^x(t_1) \prod_{j=1}^{u} \tilde{P}_2^x(t_2) \quad (31)
= \sum_{(x, t_1) \in \Lambda^n_2} \sum_{i=1}^{s} \tilde{P}_1(x) \prod_{r=1}^{s} \tilde{P}_1^x(t_1) \nu(i)
= \sum_{(q_x, q_{t_1}) \in \Gamma(\Lambda^n_2)} \sum_{i=1}^{s} \tilde{P}_1^x(T_n(q_x)) \prod_{r=1}^{s} \tilde{P}_1^x(T_n(q_{t_1})) \nu(i)
\leq \sum_{(q_x, q_{t_1}) \in \Gamma(\Lambda^n_2)} \sum_{i=1}^{s} \exp \left\{-nD(q_x||\tilde{P}_1^x) - \sum_{r=1}^{s} ND(q_{t_1}||\tilde{P}_1^x) \right\} \quad (32)
\leq \sum_{(q_x, q_{t_1}) \in \Gamma(\Lambda^n_2)} \sum_{i=1}^{s} \exp \left\{-nD(q_x||\tilde{P}_1^x) - ND(q_{t_1}||\tilde{P}_1^x) \right\}
= \sum_{(q_x, q_{t_1}) \in \Gamma(\Lambda^n_2)} \sum_{i=1}^{s} \exp \left\{-nD(q_x||q_{t_1}) - ND(q_{t_1}||q_{t_1}) - n(1 + \gamma)D \left( \frac{q_x + \gamma q_{t_1}}{1 + \gamma} \right) \right\}, \quad (33)
$$

where (31) is derived from (4), (32) follows from Lemma 1 and (33) is derived in Appendix A. Minimizing exponents in (33) with respect to $i \in \mathcal{S}$, we further obtain...
\[ \beta_1(\Lambda^n|\bar{P}_1, \bar{P}_2) \leq \sum_{(q_x,q_{t_1}) \in \Gamma(\Lambda_n^x)} s \exp \left\{ -n \min_{i \in S} \left[ D(q_x||q_{t_1}'') + \gamma D(q_{t_1}'||q_{t_1}'') \right] \right\} \cdot \exp \left\{ -n \min_{i \in S} (1 + \gamma) D \left( \frac{q_x + \gamma q_{t_1}'}{1 + \gamma} \bigg|\bigg| \bar{P}_1 \right\} \right\} \leq \sum_{(q_x,q_{t_1}) \in \Gamma(\Lambda_n^x)} s \exp \left\{ -n \bar{\lambda} \right\} \cdot \exp \left\{ -n \min_{i \in S} (1 + \gamma) D \left( \frac{q_x + \gamma q_{t_1}'}{1 + \gamma} \bigg|\bigg| \bar{P}_1 \right\} \right\} \right\} \leq \exp \left\{ -n \left[ \bar{\lambda} - \frac{\log s}{n} \right] \right\} \cdot \sum_{Q \in P_{n+N}(X)} \exp \left\{ \min_{i \in S} - (n + N) D(Q||\bar{P}_1) \right\} \right\} \leq \exp \left\{ -n \left[ \bar{\lambda} - \frac{\log s}{n} \right] \right\} \cdot (n + N + 1)^{|X|} \cdot \sum_{Q \in P_{n+N}(X)} \max_{i \in S} \bar{P}_1(i(T_{n+N}(Q))) \leq \exp \left\{ -n \left[ \bar{\lambda} - \frac{\log s}{n} \right] \right\} \cdot (n + N + 1)^{|X|} \cdot \sum_{Q \in P_{n+N}(X)} \sum_{i=1}^s \bar{P}_1(i(T_{n+N}(Q))) = \exp \left\{ -n \left[ \bar{\lambda} - \frac{2 \log s \gamma}{n} \right] \right\} \cdot \sum_{Q \in P_{n+N}(X)} \sum_{i=1}^s \bar{P}_1(i(T_{n+N}(Q))) = \exp \left\{ -n \left[ \bar{\lambda} - \frac{2 \log s \gamma}{n} \right] \right\} \cdot \sum_{i=1}^s \bar{P}_1(i(T_{n+N}(Q))) = \exp \left\{ -n \left[ \bar{\lambda} - \frac{2 \log s \gamma}{n} \right] \right\} \cdot \sum_{i=1}^s \bar{P}_1(i(T_{n+N}(Q))) = \exp \left\{ -n \left[ \bar{\lambda} - \frac{2 \log s \gamma}{n} \right] \right\} \right\} \right\} \leq \exp \left\{-n \lambda \right\}, \] 

where (34) is derived from (16) and (28), and (35) follows from Lemma 1.

Next, we demonstrate (26). For preliminaries, we define a new typical set and show some properties for the proof. For any given \( Q \in P(X) \), define the following typical set:

\[ B_n(Q) := \left\{ x \in X^n : \max_{x \in X} |q_x(x) - Q(x)| \leq \sqrt{\frac{\log n}{n}} \right\}. \] 

By using [6, Lemma 22], for \( t_1 \sim Q_1, x \sim Q_2 \), generated from memoryless sources \( (Q_1, Q_2) \in P(X)^2 \), we have

\[ \Pr \{ t_1 \notin B_N(Q_1) \text{ or } x \notin B_n(Q_2) \} \leq \frac{(1 + \gamma^2) |X|}{\gamma^2 n^2}. \] 

For any given \( x \in X \) and any pair of distributions \( (Q_1, Q_2) \in P(X)^2 \), we define information density as

\[ \iota_i(x|Q_1, Q_2, \gamma) := \log \frac{(1 + \gamma) Q_i(x)}{\gamma Q_1(x) + Q_2(x)} \quad (i = 1, 2). \]
Furthermore, for any pair of distributions \((Q_1, Q_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\), define the function \(i^*(Q_1, Q_2)\), as the index of the subclass of \(Q_1\), as follows:

\[
i^*(Q_1, Q_2) := \arg \min_{i \in S} \left\{ \gamma D \left( Q_i \left\| \frac{\gamma Q_1 + Q_2}{1 + \gamma} \right\| \right) + D \left( Q_2 \left\| \frac{\gamma Q_1 + Q_2}{1 + \gamma} \right\| \right) \right\}
= \arg \min_{i \in S} \text{GJS}(Q_i, Q_2).\tag{40}
\]

Hereafter, we denote \(i^*(Q_1, Q_2)\) simply as \(i^*(Q_2)\) when the first argument is clear from the context.

**Lemma 2.** Assume that \((Q_1, Q_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\). For \(q_{t_i} \in \mathcal{B}_N(Q_i)\) \((i \in S)\), \(q_x \in \mathcal{B}_n(Q_2)\), we have

\[
i^*(q_{t_i}, q_x) \to i^*(Q_1, Q_2) \quad (n \to \infty).\tag{41}
\]

**Proof.** The proof is provided in Appendix B. \(\square\)

**Lemma 3** (Zhou et al. [5]). Assume that \((Q_1, Q_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\). For \(q_{t_i} \in \mathcal{B}_N(Q_i)\) \((i \in S)\), \(q_x \in \mathcal{B}_n(Q_2)\), by applying the Taylor expansion to \(\text{GJS}(q_{t_i}, q_x)\) around \((Q_i, Q_2)\), we have

\[
\text{GJS}(q_{t_i}, q_x) = \gamma \frac{N}{n} \sum_{m=1}^{n} t_1(m)Q_1(t_1, Q_2, \gamma) + \frac{1}{n} \sum_{m=1}^{n} t_2(x_m)Q_1(Q_2, \gamma) + O \left( \frac{\log n}{n} \right), \tag{42}
\]

where \(t_1(m)\) denotes the \(m\)-th symbol of sequence \(t_1\) and \(x_m\) denotes the \(m\)-th symbol of sequence \(x\).

Note that the probability \(P_2\) is calculated by assuming that the test sequence \(x\) is generated from \(P_2\) (cf. Equation (11)). An upper bound on the type-II error probability can be evaluated as follows:

\[
\limsup_{n \to \infty} \beta_2(A^n | P_1, P_2) \\
= \limsup_{n \to \infty} P_2 \left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) < \lambda + \eta(n) \right\} \\
= \limsup_{n \to \infty} \sum_{j=1}^{n} P_j \left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) < \lambda + \eta(n) \right\} \\
\leq \limsup_{n \to \infty} \sum_{j=1}^{n} P_j \left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) < \lambda + \eta(n), x \in \mathcal{B}_n(P_j), t_1 \in \mathcal{B}_n(P_j), \forall i \in S \right\} \\
\quad + \limsup_{n \to \infty} \sum_{j=1}^{n} P_j \left\{ x \notin \mathcal{B}_n(P_j) \text{ or } t_1 \notin \mathcal{B}_n(P_j) \right\} \\
\leq \liminf_{n \to \infty} \sum_{j=1}^{n} P_j \left\{ \text{GJS}(q_{t_1}, q_x, \gamma) < \lambda + \eta(n), x \in \mathcal{B}_n(P_j), t_1 \in \mathcal{B}_n(P_j), \forall i \in S \right\}, \tag{43}
\]

where Equation (43) follows from Equations (14), (38) and (40). By using Lemma 3, Equation (43) can be expanded as follows:
\[
\limsup_{n \to \infty} \beta_2(\Lambda^n | P_1, P_2) \\
\leq \limsup_{n \to \infty} \sum_{j=1}^{\infty} w(j) \mathbb{P}_2^j \left\{ \frac{\gamma}{N} \sum_{m=1}^{N} t_1(i^{\ast}_{1,m} | P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) + \frac{1}{n} \sum_{m=1}^{n} t_2(x_m | P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) < \lambda \right\} \\
+ O\left(\frac{\log n}{n}\right), x \in B_n(P_2^j), t_1^j \in B_N(P_1^j), \forall i \in S \\
\leq \sum_{j=1}^{\infty} \limsup_{n \to \infty} w(j) \mathbb{P}_2^j \left\{ \frac{\gamma}{N} \sum_{m=1}^{N} t_1(i^{\ast}_{1,m} | P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) + \frac{1}{n} \sum_{m=1}^{n} t_2(x_m | P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) < \lambda \right\} \\
+ O\left(\frac{\log n}{n}\right), x \in B_n(P_2^j), t_1^j \in B_N(P_1^j), \forall i \in S \\
\leq \sum_{j=1}^{\infty} \limsup_{n \to \infty} w(j) \mathbb{P}_2^j \left\{ x \in B_n(P_2^j), t_1^j \in B_N(P_1^j), \forall i \in S \right\} \\
+ O\left(\frac{\log n}{n}\right) \left\{ x \in B_n(P_2^j), t_1^j \in B_N(P_1^j), \forall i \in S \right\} \\
= \sum_{j=1}^{\infty} \limsup_{n \to \infty} w(j) \mathbb{P}_2^j \left\{ x \in B_n(P_2^j), t_1^j \in B_N(P_1^j), \forall i \in S \right\} \\
+ O\left(\frac{\log n}{n}\right) \left\{ x \in B_n(P_2^j), t_1^j \in B_N(P_1^j), \forall i \in S \right\},
\]

where Equation (44) is derived from Fatou’s lemma. It follows from (38) that
\[
\mathbb{P}_2^j \left\{ t_1^j \in B_n(P_1^j), x \in B_n(P_2^j), \forall i \in S \right\} \to 1 \quad (n \to \infty).
\] (46)

Here, Equation (14) can be also expressed as follows:
\[
\text{GJS}(Q_1, Q_2, \gamma) = \gamma \mathbb{E}_{Q_1} [t_1(X|Q_1, Q_2, \gamma)] + \mathbb{E}_{Q_2} [t_2(X|Q_1, Q_2, \gamma)].
\] (47)

Therefore, by the weak law of large numbers, for any given \( \delta > 0 \),
\[
\limsup_{n \to \infty} \text{Pr} \left\{ \left| \frac{\gamma}{N} \sum_{m=1}^{N} t_1(i^{\ast}_{1,m} | P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) + \frac{1}{n} \sum_{m=1}^{n} t_2(x_m | P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) - \text{GJS}(P_1^\ast(q_x), P_2^\ast(q_x), P_1^j, P_2^j, \gamma) \right| \leq \delta \right\} = 1.
\] (48)

From this result of the weak law of large numbers and Lemma 2, combining (45)–(48) gives
\[
\limsup_{n \to \infty} \beta_2(\Lambda^n | P_1, P_2) \leq \sum_{\{j \in \mathcal{U} : \text{GJS}(P_1^j, P_2^j) < \lambda\}} w(j).
\] (49)

Thus, by (24), we can see that
\[
\limsup_{n \to \infty} \beta_2(\Lambda^n | P_1, P_2) \leq \epsilon.
\] (50)
3.3.2. Converse Part

For any pair of distributions \((P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\) and for all pairs of distributions \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\), fix any test \(\phi^n\) satisfying that

\[
\beta_1(\phi^n|\tilde{P}_1, \tilde{P}_2) \leq \exp(-n\lambda),
\]

\[
\limsup_{n \to \infty} \beta_2(\phi^n|P_1, P_2) \leq \epsilon.
\]

We show

\[
\lambda \leq \sup \left\{ \lambda \left| \sum_{j \in U : MGJS(P_1, P_2, \gamma) < \lambda} w(j) \leq \epsilon \right. \right\}.
\]

We first give some lemmas, which are useful in the proof of the converse part.

**Lemma 4.** Let \(\phi^n\) be a test in which the decision rule depends only on \((x, t_1, t_2) \in \mathcal{X}^n \times \mathcal{X}^s \times \mathcal{X}^u N\). Then, for any given \(\kappa \in [0, 1]\), we can construct a type-based test \(\Omega^n\) satisfying

\[
\beta_1(\phi^n|\tilde{P}_1, \tilde{P}_2) \geq \kappa \beta_1(\Omega^n|\tilde{P}_1, \tilde{P}_2),
\]

\[
\beta_2(\phi^n|\tilde{P}_1, \tilde{P}_2) \geq (1 - \kappa) \beta_2(\Omega^n|\tilde{P}_1, \tilde{P}_2)
\]

for any pair of distributions \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\).

**Proof.** Lemma 4 can be proved in the same way as (Lemma 7 [5]), the proof of which is inspired by (Lemma 2 [3]). \(\square\)

**Remark 3.** As in the proof of (Lemma 2 [3]) and (Lemma 7 [5]), a type-based test \(\Omega^n\) specified in Lemma 4 is obtained by tailoring \(\phi^n\) and satisfies Equations (54) and (55) for all \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\). In other words, the construction of \(\Omega^n\) is universal, which is in the same spirit of (Lemma 2 [3]). This claim is slightly stronger than the one in (Lemma 7 [5]).

**Lemma 5.** For any \(\lambda \in \mathbb{R}^+\), any type-based test \(\Omega^n\) satisfying the condition that for all pairs of distributions \((P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\),

\[
\beta_1(\Omega^n|P_1, P_2) \leq \exp(-n\lambda),
\]

we have that for any pair of distributions \((P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\)

\[
\beta_2(\Omega^n|P_1, P_2) \geq P_2\{|MGJS(q_1, q_2, \gamma) < \lambda - \rho(n)\},
\]

where \(\rho(n) := \frac{|\mathcal{X}| |\log(n + 1) + (s + u) | \log(N + 1) - \log n^*}{n}\) with \(n^* := \min\{v(i) : v(i) > 0, i \in \mathcal{S}\}\).

**Proof.** The proof is provided in Appendix C. \(\square\)

The type-based test \(\Omega^n\) specified in Lemma 4 satisfies Equations (54) and (55) for all \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\). If we set \(\kappa = 1/n\) in Lemma 4, and combine it with Lemma 5, we can derive the following relation:

**Corollary 1.** For any given \(\lambda \in \mathbb{R}^+\), any test \(\phi^n\) satisfying the condition that for all pairs of distributions \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\)

\[
\beta_1(\phi^n|\tilde{P}_1, \tilde{P}_2) \leq \exp(-n\lambda),
\]
we have that for any pair of distributions \((P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\)

\[
\beta_2(p^n|P_1, P_2) \geq \left(1 - \frac{1}{n}\right) \cdot P_2\left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) + o(n) + \frac{\log n}{n} < \lambda \right\}.
\]  

(59)

**Proof.** The proof is provided in Appendix D. \(\square\)

By (59), a lower bound on the type-II error probability can be evaluated as follows:

\[
\limsup_{n \to \infty} \beta_2(p^n|P_1, P_2) \\
\geq \limsup_{n \to \infty} \left(1 - \frac{1}{n}\right) P_2\left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) + O\left(\frac{\log n}{n}\right) < \lambda \right\} \\
= \limsup_{n \to \infty} \sum_{j=1}^u w(j) P_2\left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) + O\left(\frac{\log n}{n}\right) < \lambda, x \in B_n(p_j), t_1 \in B_N(P_i), \forall i \in S \right\} \\
\geq \limsup_{n \to \infty} \sum_{j=1}^u w(j) P_2\left\{ \frac{1}{n} \sum_{m=1}^n t_1(f_{t_1,m}^{(q_x)}|p_1^{(q_x)}, p_{j_1}^{(q_x)}), p_{j_1}^{(q_x)}, \gamma) + \frac{1}{n} \sum_{m=1}^n t_2(x_m|p_{1}^{(q_x)}, p_{2}^{(q_x)}) \right. \\
+ O\left(\frac{\log n}{n}\right) < \lambda, x \in B_n(p_j), t_1 \in B_N(P_i), \forall i \in S \right\} \\
\geq \sum_{j=1}^u \liminf_{n \to \infty} w(j) P_2\left\{ x \in B_n(p_j), t_1 \in B_N(P_i), \forall i \in S \right\} \\
\cdot P_2\left\{ \frac{\gamma}{N} \sum_{m=1}^N t_1(f_{t_1,m}^{(q_x)}|p_1^{(q_x)}, p_{j_1}^{(q_x)}), p_{j_1}^{(q_x)}, \gamma) + \frac{1}{n} \sum_{m=1}^n t_2(x_m|p_{1}^{(q_x)}, p_{2}^{(q_x)}) < \lambda \right. \\
+ O\left(\frac{\log n}{n}\right) \left| x \in B_n(p_j), t_1 \in B_N(P_i), \forall i \in S \right\} \\
= \sum_{\{j \in \mathcal{U}: \text{MGJS}(p_j^{(q_x)}, p_{j_1}^{(q_x)}, \gamma) < \lambda\}} w(j),
\]

(60)

(61)

(62)

where Equations (60) and (61) are derived from Lemma 3 and Fatou’s lemma, respectively. Equation (62) follows from Equations (45)–(48). By Equation (52), Equation (62) indicates that

\[
\lambda \leq \sup \left\{ \lambda \left| \sum_{\{j \in \mathcal{U}: \text{MGJS}(p_j^{(q_x)}, p_{j_1}^{(q_x)}, \gamma) < \lambda\}} w(j) \leq \epsilon \right\} \right. \\
\]

(63)

3.4. Second-Order Maximum Error Exponent

In this section, we characterize the second-order maximum error exponent. For simplicity, we assume that only \(P_2\) has subclasses, but \(P_1\) does not \((s = 1)\). First, from Theorem 2
with \( s = 1 \), the first-order maximum error exponent in this setting is characterized as follows: for any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\), we have

\[
\lambda(\epsilon) = \sup \left\{ \lambda \left| \sum_{j \in \mathcal{U} : GJS(P_1, P_2, \gamma) < \lambda} w(j) \leq \epsilon \right. \right\},
\]

(64)

where MGJS\((P_1, P_2^1, \gamma)\) in (21) is replaced by GJS\((P_1, P_2^1, \gamma)\).

Next, we provide a characterization of the second-order maximum error exponent in Definition 2 with \( s = 1 \) in the case where only \( P_2 \) has subclasses. By definition, \( \hat{r}(\epsilon, \lambda) = +\infty \) if \( \lambda < \hat{\lambda}(\epsilon) \) and \( \hat{r}(\epsilon, \lambda) = -\infty \) if \( \lambda > \hat{\lambda}(\epsilon) \). Therefore, in the discussion of the second-order error exponent, we focus on the case \( \lambda = \hat{\lambda}(\epsilon) \).

**Theorem 3.** For any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\) and \( \epsilon \in [0, 1) \),

\[
\hat{r}(\epsilon, \lambda) = \sup \left\{ \lambda \left| \sum_{j \in \mathcal{U} : GJS(P_1, P_2^1, \gamma) < \lambda} w(j) + \sum_{j \in \mathcal{U} : GJS(P_1, P_2^1, \gamma) = \lambda} \Phi_j(r) w(j) \leq \epsilon \right. \right\},
\]

(65)

where \( \lambda = \hat{\lambda}(\epsilon) \),

\[
\Phi_j(r) := G\left( \frac{r}{\sqrt{V(P_1, P_2^1, \gamma)}} \right),
\]

(66)

\[
G(a) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx,
\]

(67)

which is the cumulative distribution function of the standard Gaussian distribution, and for any pair of distributions \((Q_1, Q_2) \in \mathcal{P}(\mathcal{X})^2\),

\[
V(Q_1, Q_2, \gamma) := \gamma \text{Var}_{Q_1}[1(X|Q_1, Q_2, \gamma)] + \text{Var}_{Q_2}[2(X|Q_1, Q_2, \gamma)],
\]

(68)

where \( \text{Var}_{Q}[\cdot] \) represents the variances with respect to \( Q \in \mathcal{P}(\mathcal{X}) \).

**Proof.** The proof is provided in Appendix E. \( \square \)

**Remark 4.** If \( \mathcal{U} \) is a singleton \((u = 1)\), Theorem 3 reduces to \( \hat{r}(\epsilon, \lambda) = \sqrt{V(P_1, P_2^1, \gamma)}^{-1}(\epsilon) \) for \( \lambda = \hat{\lambda}(\epsilon) = \text{GJS}(P_1, P_2^1, \gamma) \), which is the same result given by Zhou et al. [5].

**Remark 5.** We can summarize the two terms on the right-hand side of (65) into the following single term called the canonical equation [6]:

\[
\sum_{j \in \mathcal{U}} w(j) \lim \sqrt{n}(\lambda - \text{GJS}(P_1, P_2^1, \gamma)) + r).
\]

(69)

We focus on the case \( \lambda = \hat{\lambda}(\epsilon) \). From Theorem 2, it holds that

\[
\sum_{\{j \in \mathcal{U} : \text{MGJS}(P_1, P_2^1, \gamma) < \lambda\}} w(j) \leq \epsilon
\]

(70)

and

\[
\sum_{\{j \in \mathcal{U} : \text{MGJS}(P_1, P_2^1, \gamma) \leq \lambda\}} w(j) \geq \epsilon.
\]

(71)
Here, let us consider the following canonical equation for $r$

$$
\sum_{j \in I} w(j) \lim_{\epsilon \to \infty} \Phi_j\left(\sqrt{\epsilon}(\hat{\lambda}(\epsilon) - \text{GJS}(P_1, P_2^\epsilon, \gamma)) + r\right) = \epsilon. \tag{72}
$$

Thus, in view of (70) and (71), this equation always has the solution $r = r(\epsilon)$. If

$$
\sum_{\{j \in I: \text{GJS}(P_1, P_2^\epsilon, \gamma) = \lambda(\epsilon)\}} w(j) = 0, \tag{73}
$$

the solution is not unique ($r(\epsilon) = +\infty$). By using the solution $r(\epsilon)$, Equation (65) with $\lambda = \hat{\lambda}(\epsilon)$ can be expressed in a simpler form

$$
\hat{r}(\epsilon, \lambda) = r(\epsilon). \tag{74}
$$

4. Generalization to Mixed Memoryless Sources with General Mixture

In this section, we consider the classification problem in the case where $P_1$ does not have subclasses and $P_2$ is given by a general mixture model. The general mixture model considered in this problem represents an extension of the source with multiple subclasses defined in Section 2.2. Since the decision rule that achieves the maximum error exponent can be operated using only one of the training sequences, we assume in this section that only the training sequence $t_1$ is available. Then, we provide a characterization of the maximum error exponents in a single-letter form under this setting. First, we define the source referred to as a mixed memoryless source with general mixture [6,9] as follows. Let $\Theta$ be an arbitrary probability space with a general probability measure $\text{w}(\theta), \theta \in \Theta$. Then, the probability of $x \in X^n$ is given by

$$
P_2(x) = \int_{\theta \in \Theta} P_2^\theta(x) d\text{w}(\theta), \tag{75}
$$

where $P_2^\theta$ is a stationary and memoryless source for each $\theta \in \Theta$. That is, for $x = (x_1, x_2, \ldots, x_n) \in X^n$

$$
P_2^\theta(x) = \prod_{i=1}^n P_2^\theta(x_i). \tag{76}
$$

When a test sequence is output from $P_2$, the probability distribution of the sequence takes the form of (75). Here, type-I and type-II error probabilities of a test $\phi^n = (\phi_1^n, \phi_2^n)$ are given by

$$
\beta_1(\phi^n | P_1, P_2) = \sum_{(x, t_1) \in \phi_1^n} P_1(x) P_1(t_1), \tag{77}
$$

$$
\beta_2(\phi^n | P_1, P_2) = \sum_{(x, t_1) \in \phi_2^n} P_2(x) P_1(t_1). \tag{78}
$$

Theorem 4. (First- and second-order maximum error exponents) For any pair of distributions $(P_1, P_2) \in \mathcal{P}(X) \times \mathcal{P}_\Theta(X)$ and $\epsilon \in [0, 1)$, we have

$$
\hat{\lambda}(\epsilon) = \sup \left\{ \lambda \left| \int_{\theta \in \Theta: \text{GJS}(P_1, P_2^\epsilon, \gamma) < \lambda} d\text{w}(\theta) \leq \epsilon \right. \right\}, \tag{79}
$$

and

$$
\hat{r}(\epsilon, \lambda) = \sup \left\{ r \left| \int_{\theta \in \Theta: \text{GJS}(P_1, P_2^\epsilon, \gamma) < \lambda} d\text{w}(\theta) + \int_{\theta \in \Theta: \text{GJS}(P_1, P_2^\epsilon, \gamma) = \lambda} \Phi_\theta(r) d\text{w}(\theta) \leq \epsilon \right. \right\}. \tag{80}
$$
Proof. We can prove this theorem in the same way as Theorems 2 and 3.

5. Numerical Calculation
5.1. First-Order Maximum Error Exponent

In this section, we present a numerical example to illustrate the first-order maximum type-I error exponent $\hat{\lambda}(\epsilon)$ characterized in Theorem 2.

A numerical example of the first-order maximum error exponent is given by calculating the right-hand side of (64) for the following settings. We assume that $\mathcal{X} = \{0, 1\}$. We fix the set of probabilities and weights

$$P_1 = \left\{ \left\{ \text{Bern}(0.389), v(1) = \frac{1}{3} \right\}, \left\{ \text{Bern}(0.322), v(2) = \frac{1}{3} \right\}, \left\{ \text{Bern}(0.256), v(3) = \frac{1}{3} \right\} \right\}$$

and

$$P_2 = \left\{ \left\{ \text{Bern}(0.301), w(1) = \frac{1}{3} \right\}, \left\{ \text{Bern}(0.244), w(2) = \frac{1}{6} \right\}, \left\{ \text{Bern}(0.223), w(3) = \frac{1}{2} \right\} \right\},$$

where $\text{Bern}(\cdot)$ denotes the Bernoulli distribution. The relation among $\hat{\lambda}(\epsilon)$, $\epsilon$ and $\gamma$ is shown in Figure 2. Additionally, for $\gamma = 2$, the behavior of $\hat{\lambda}(\epsilon)$ is depicted in Figure 3. When $\epsilon$ becomes larger, the value of $\hat{\lambda}(\epsilon)$ also increases like a step function. The step increases when $\epsilon = \frac{1}{6}$ and $\epsilon = \frac{1}{2}$. We can also confirm that $\hat{\lambda}(\epsilon)$ is right-continuous in $\epsilon$. This is due to the limit that the superior of the type-II error probability is constrained in Definition 1.

![Figure 2](image_url)

**Figure 2.** The first-order maximum type-I error exponent $\hat{\lambda}(\epsilon)$ ($0 < \gamma \leq 2$).

![Figure 3](image_url)

**Figure 3.** The first-order maximum type-I error exponent $\hat{\lambda}(\epsilon)$ ($\gamma = 2$).
5.2. Second-Order Maximum Error Exponent

As in the previous subsection, we present a numerical example to illustrate the second-order maximum type-I error exponent \( \hat{r}(\epsilon, \lambda) \) characterized in Theorem 3.

A numerical example of the second-order maximum error exponent is given by calculating the right-hand side of (65) for the following settings. We assume that \( \gamma = 2, \mathcal{X} = \{0, 1\} \). We fix \( \lambda = \hat{\lambda}(\epsilon) \) and the set of probabilities and weights

\[
P_1 = \text{Bern}(0.268) \tag{83}
\]

and

\[
P_2 = \left\{ \left\{ \text{Bern}(0.301), w(1) = \frac{1}{3} \right\}, \left\{ \text{Bern}(0.244), w(2) = \frac{1}{6} \right\}, \left\{ \text{Bern}(0.223), w(3) = \frac{1}{2} \right\} \right\}, \tag{84}
\]

where \( P_2 \) is the same as the setting in the previous subsection. The behavior of \( \hat{r}(\epsilon, \lambda) \) is shown in Figure 4. The value of \( \hat{r}(\epsilon, \lambda) \) takes the inverse of the cumulative distribution function of the standard Gaussian for each interval of \( \epsilon \) such that \( 0 \leq \epsilon < \frac{1}{6}, \frac{1}{6} \leq \epsilon < \frac{1}{2} \), and \( \frac{1}{2} \leq \epsilon < 1 \). In contrast to the first-order \( \hat{\lambda}(\epsilon) \), \( \hat{r}(\epsilon, \lambda) \) is no longer right continuous in \( \epsilon \).

6. Conclusions

For binary classification of sources with multiple subclasses, we characterized the first- and second-order maximum error exponents. First, we revealed the first-order maximum error exponent in the case where \( P_1 \) and \( P_2 \) are sources with multiple subclasses. In order to derive this representation, we gave a classifier which achieves the asymptotically maximum error exponent in the class of deterministic classifiers for sources with multiple subclasses.

Next, we showed the second-order maximum error exponent in the case where only one of sources has subclasses. The most important key technique to derive the second-order maximum error exponent is to apply the Berry–Esseen theorem [8] instead of the weak law of large numbers. One may wonder whether we can also derive the second-order approximation in the case where \( P_1 \) is also a source with multiple subclasses. To this end, we need to evaluate Lemma 2 more rigorously. This is future work.

In addition, for binary classification using only a training sequence generated from \( P_1 \) in the case where \( P_1 \) does not have subclasses and \( P_2 \) is given by a general mixture model, we generalized the analysis for the first- and second-order error exponents. From these results, we revealed the asymptotic performance limits of statistical classification for sources with multiple subclasses.

In this paper, we considered a binary classification problem, but in practice, multiclass classification is of importance. In the case where each class is a memoryless source (without multiple subclasses), the first- and second-order maximum error exponents were analyzed.
in [5]. Extending the obtained results to multiclass classification for sources with multiple subclasses is also a subject of future studies.

Author Contributions: Author H.K. contributed to the conceptualization of the research goals and aims, the visualization, the formal analysis of the results, and the review and editing. Author H.Y. contributed to the conceptualization of the ideas, the validation of the results, and the supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by JSPS KAKENHI Grant Numbers JP20K04462 and JP18H01438.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Equation (33). We shall show

\[ ND(q_t \mid | \tilde{P}_1) + nD(q_x \mid | \tilde{P}_1) = ND(q_t \mid | q_y) + nD(q_x \mid | q_y) + n(1 + \gamma)D \left( \frac{\gamma q_t + q_x}{1 + \gamma} \mid | P_1 \right). \]  

(A1)

The left-hand side of (A1) is expanded as follows:

\[
ND(q_t \mid | \tilde{P}_1) + nD(q_x \mid | \tilde{P}_1) \\
= n\gamma \sum_{x \in \mathcal{X}} q_t(x) \log \frac{q_t(x)}{\tilde{P}_1(x)} + n \sum_{x \in \mathcal{X}} q_x(x) \log \frac{q_x(x)}{\tilde{P}_1(x)} - n\gamma \sum_{x \in \mathcal{X}} q_t(x) \log \frac{\gamma q_t(x) + q_x(x)}{1 + \gamma} \\
+ n \sum_{x \in \mathcal{X}} q_x(x) \log \frac{\gamma q_t(x) + q_x(x)}{1 + \gamma} \\
= n\gamma \mathbb{E}_{q_t} \left[ \log \frac{1 + \gamma q_t(x)}{\gamma q_t(x) + q_x(x)} \right] + n\mathbb{E}_{q_x} \left[ \log \frac{(1 + \gamma)q_x(x)}{\gamma q_t(x) + q_x(x)} \right] \\
- n \sum_{x \in \mathcal{X}} (\gamma q_t(x) + q_x(x)) \log \tilde{P}_1(x) + n \sum_{x \in \mathcal{X}} (\gamma q_t(x) + q_x(x)) \log \frac{\gamma q_t(x) + q_x(x)}{1 + \gamma} \\
= ND(q_t \mid | q_y) + nD(q_x \mid | q_y) + n(1 + \gamma)D \left( \frac{\gamma q_t + q_x}{1 + \gamma} \mid | P_1 \right). 
\]  

(A2)

Therefore, we obtain Equation (33). □

Appendix B

Proof of Lemma 2. Applying the Taylor expansion to GJS($q_{t_1}$, $q_x$, $\gamma$) around ($Q_1$, $Q_2$) for any $q_x \in B_n(Q_2), q_{t_1} \in B_n(Q_1)$ ($\forall i \in S$), we obtain

\[
GJS(q_{t_1}, q_x, \gamma) = GJS(Q_1, Q_2, \gamma) + O \left( \sqrt{\frac{\log n}{n}} \right). 
\]

(A3)
Therefore by (14) for any \( q_x \in \mathcal{B}_n(Q_2), q_{t_1} \in \mathcal{B}_N(Q_1) \) (\( \forall i \in S \)),

\[
\gamma D \left( q_{t_1} \mid \mid \frac{q_x + \gamma q_{t_1}}{1 + \gamma} \right) + D \left( q_x \mid \mid \frac{q_x + \gamma q_{t_1}}{1 + \gamma} \right) \\
\rightarrow \gamma D \left( Q_1 \mid \mid \frac{Q_2 + \gamma Q_1}{1 + \gamma} \right) + D \left( Q_2 \mid \mid \frac{Q_2 + \gamma Q_1}{1 + \gamma} \right) \quad (n \to \infty)
\]

(A4)

holds, and the convergence is uniform in \( i \in S \). Since for any \( (Q_1, Q_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \), \( i^*(Q_1, Q_2) \) was given in the form

\[
i^*(Q_1, Q_2) = \arg \min_{i \in S} \gamma D \left( Q_1 \mid \mid \frac{Q_2 + \gamma Q_1}{1 + \gamma} \right) + D \left( Q_2 \mid \mid \frac{Q_2 + \gamma Q_1}{1 + \gamma} \right),
\]

(A5)

we obtain

\[
i^* (q_{t_1}, q_x) \to i^*(Q_1, Q_2) \quad (n \to \infty).
\]

(A6)

\( \square \)

**Appendix C**

**Proof of Lemma 5.** It follows from (56) that

\[
\exp \{ -n \lambda \} \\
\geq \sum_{(x, t_1, t_2) \in \Omega^n_2} \tilde{P}_1(x) \prod_{r=1}^{s} \tilde{P}_1^r(t_1^r) \prod_{j=1}^{u} \tilde{P}_2^j(t_2^j) \\
= \sum_{(x, t_1, t_2) \in \Omega^n_2} \sum_i^{\infty} \tilde{P}_1(x) \prod_{r=1}^{s} \tilde{P}_1^r(t_1^r) \prod_{j=1}^{u} \tilde{P}_2^j(t_2^j) \nu(i) \\
\geq \sum_{(x, t_1, t_2) \in \Omega^n_2} \tilde{P}_1(x) \prod_{r=1}^{s} \tilde{P}_1^r(T_N(q_{x})) \prod_{j=1}^{u} \tilde{P}_2^j(T_N(q_{t_2})) \nu(i) \\
\geq \sum_{(x, t_1, t_2) \in \Omega^n_2} \tilde{P}_1(x) \prod_{r=1}^{s} \tilde{P}_1^r(T_N(q_{x})) \prod_{j=1}^{u} \tilde{P}_2^j(T_N(q_{t_2})) \nu(i) \\
\quad \forall (q_{x}, q_{t_1}, q_{t_2}) \in \Gamma(\Omega^n_2).
\]

(A7)

Using (30) on the right-hand side of (A7), we obtain

\[
\exp \{ -n \lambda \} \\
\geq \exp \{ -n \rho(n) \} \exp \left\{ -n \left[ D(q_x || \tilde{P}_1^r) + \sum_{r=1}^{s} \gamma D(q_{t_1}^r || \tilde{P}_1^r) + \sum_{j=1}^{u} \gamma D(q_{t_2}^r || \tilde{P}_2^j) \right] \right\},
\]

(A8)

where \( \rho(n) = \frac{|X| \log(n+1) + (s+u)|X| \log(N+1) - \log \nu}{n} \). Taking the negative logarithm of both sides and divide by \( n \) in (A8), we have

\[
\lambda - \rho(n) \leq D(q_x || \tilde{P}_1^r) + \sum_{r=1}^{s} \gamma D(q_{t_1}^r || \tilde{P}_1^r) + \sum_{j=1}^{u} \gamma D(q_{t_2}^r || \tilde{P}_2^j).
\]

(A9)
Since \((\tilde{P}_1, \tilde{P}_2, \tilde{P}_3) \in \mathcal{P}(\mathcal{X})^3\) is arbitrary in (A9), we can set the \(i\)-th subclass as 
\[
\tilde{P}_i = q_{xti} = q_{yi},
\]
and the other as \(\tilde{P}_r = q_{tr}, r \neq i\). Furthermore, we set \(\tilde{P}_j = q_{tj}, \forall j \in \mathcal{U}\). Then we obtain
\[
\lambda - \rho(n) = D(q_x || q_{y_i}) + \gamma D(q_{t_1} || q_{y_i}) \quad (\forall (q_x, q_{t_1}, q_{t_2}) \in \Gamma(\Omega_\beta^i), \forall i \in \mathcal{S})
\]
\[
\iff \lambda - \rho(n) = \min_{i \in \mathcal{S}} \left\{ D(q_x || q_{y_i}) + \gamma D(q_{t_1} || q_{y_i}) \right\} \quad (\forall (q_x, q_{t_1}, q_{t_2}) \in \Gamma(\Omega_\beta^i)). \tag{A10}
\]
Therefore by (18), (A10) implies that \(\Omega_\beta^i\) is constrained in
\[
\bar{\Lambda}_\beta^i := \{(x, t_1, t_2) \mid \text{MGJS}(q_{t_1}, q_x, \gamma) \geq \lambda - \rho(n)\}, \tag{A11}
\]
and for any pair of distributions \(P = (P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\), we obtain
\[
\Omega_\beta^2 \subset \bar{\Lambda}_\beta^i \Rightarrow \bar{\Lambda}_\beta^i \subset \Omega_\beta^2 \Rightarrow \beta_2(\bar{\Lambda}_\beta^i | P) \leq \beta_2(\Omega_\beta^2 | P). \tag{A12}
\]
\[\square\]

Appendix D

Proof of Corollary 1. It follows from (58) that
\[
\exp(-n\lambda) \geq \beta_1(\phi^n | \tilde{P}_1, \tilde{P}_2)
\]
\[
\geq \frac{1}{n} \beta_1(\Omega^n | \tilde{P}_1, \tilde{P}_2), \tag{A13}
\]
where Lemma 4 with \(\kappa = 1/n\) guarantees the existence of such a type-based test \(\Omega^n\) satisfying (A13) for \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\) and (55) for \((P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\) (cf. Remark 3). Then, the type-I error probability of \(\Omega^n\) satisfies
\[
\beta_1(\Omega^n | \tilde{P}_1, \tilde{P}_2) \leq n \exp(-n\lambda)
\]
\[
= \exp\left(-n\left(\lambda - \frac{\log n}{n}\right)\right)
\]
\[
= \exp(-n\lambda'), \tag{A14}
\]
where \(\lambda' = \lambda - \frac{\log n}{n}\). From Lemma 5, the type-II error probability of \(\Omega^n\) satisfies
\[
\beta_2(\Omega^n | P_1, P_2) \geq \mathbb{P}_2\left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) < \lambda' - \rho(n) \right\}. \tag{A15}
\]
Combining (A15) and (55) for \((P_1, P_2) \in \mathcal{P}_S(\mathcal{X}) \times \mathcal{P}_U(\mathcal{X})\), we have
\[
\beta_2(\phi^n | P_1, P_2) \geq \left(1 - \frac{1}{n}\right) \mathbb{P}_2\left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) < \lambda' - \rho(n) \right\}
\]
\[
= \left(1 - \frac{1}{n}\right) \mathbb{P}_2\left\{ \text{MGJS}(q_{t_1}, q_x, \gamma) + \rho(n) + \frac{\log n}{n} < \lambda \right\}, \tag{A16}
\]
establishing (59). \(\square\)
Appendix E

Proof of Theorem 3. We divide the proof of Theorem 3 into two parts: the achievability (direct) part and the converse part. First, for preliminaries, we define the following quantity used in the proof: For \((Q_1, Q_2) \in \mathcal{P}(\mathcal{X})^2\),

\[
T(Q_1, Q_2, \gamma) := \gamma E_{Q_1} \left[ |t_1(X|Q_1, Q_2, \gamma) - E_{Q_1}[|t_1(X|Q_1, Q_2, \gamma)]|^3 \right] + E_{Q_2} \left[ |t_2(X|Q_1, Q_2, \gamma) - E_{Q_2}[|t_2(X|Q_1, Q_2, \gamma)]|^3 \right].
\]

(A17)

Next, we give a lemma that is important in the proof of Theorem 3.

Lemma A1 (The Berry–Esseen theorem [8]). Let \(X_k, k = 1, ..., n\) be independent with

\[
\mu_k = E[X_k],
\]

(A18)

\[
\sigma^2 = \sum_{k=1}^{n} \text{Var}[X_k],
\]

(A19)

\[
T = \sum_{k=1}^{n} E[|X_k - \mu_k|^3],
\]

(A20)

\[
Q(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)dt.
\]

(A21)

Then for any \(-\infty < \lambda < \infty\), it holds that

\[
\left| \mathbb{P}\left[\sum_{k=1}^{n} (X_k - \mu_k) \geq \lambda \sigma \right] - Q(\lambda) \right| \leq \frac{6T}{\sigma^3}.
\]

(A22)

Appendix E.1. Achievability Part

In the achievability proof, we use the following test \(\Lambda^n = (\Lambda^n_1, \Lambda^n_2)\):

\[
\Lambda^n_2 = \left\{(x, t_1, t_2) \mid \text{GJS}(q_{t_1}, q_{x}, \gamma) \geq \bar{\lambda} + \frac{r}{\sqrt{n}} \right\},
\]

(A23)

where \(\text{MGJS}(q_{t_1}, q_{x}, \gamma)\) in (17) is now \(\text{GJS}(q_{t_1}, q_{x}, \gamma)\) because the first source is assumed to be a stationary and memoryless source. Fix any

\[
r < \sup \left\{ r \left| \sum_{\{j \in U : \text{GJS}(P_1, P_2, \gamma) < \lambda\}} w(j) + \sum_{\{j \in U : \text{GJS}(P_1, P_2, \gamma) = \lambda\}} \Phi_j(r)w(j) \leq \epsilon \right. \right\}
\]

(A24)

Then, for any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}_d(\mathcal{X})\) and for all pairs of distributions \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}_d(\mathcal{X})\), we show

\[
\beta_1(\Lambda^n | \tilde{P}_1, \tilde{P}_2) \leq \exp(-n\bar{\lambda} - \sqrt{n}r),
\]

(A25)

\[
\limsup_{n \to \infty} \beta_2(\Lambda^n | P_1, P_2) \leq \epsilon.
\]

(A26)

The test \(\Lambda^n\) defined by (A23) is the same as the test defined by (17) with replacing \(\bar{\lambda}\) with \(\bar{\lambda} + \frac{r}{\sqrt{n}}\) and \(s\) with 1, and so (A25) can be derived from the argument in Section 3.3.1. Therefore, we will prove only (A26). We use the set \(B_n(Q), Q \in \mathcal{P}(\mathcal{X})\) defined by (37). An upper bound on the type-II error probability of the test \(\Lambda^n\), defined by (A23), for any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}_d(\mathcal{X})\) can be evaluated as follows:
\[
\limsup_{n \to \infty} \beta_2(L^n | P_1, P_2)
\]
\[
= \limsup_{n \to \infty} \mathbb{P}_2 \left\{ \text{GJS}(q_{t_1}, q_{x}, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + \eta(n) \right\}
\]
\[
= \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \mathbb{P}_2 \left\{ \text{GJS}(q_{t_1}, q_{x}, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + \eta(n) \right\}
\]
\[
\leq \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \mathbb{P}_2 \left\{ \text{GJS}(q_{t_1}, q_{x}, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + \eta(n), t_1 \in B_N(P_1), x \in B_n(P_2) \right\}
\]
\[
+ \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \mathbb{P}_2 \left\{ t_1 \notin B_N(P_1) \text{ or } x \notin B_n(P_2) \right\}
\]
\[
= \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \mathbb{P}_2 \left\{ \frac{1}{n} \sum_{i=1}^{N} t_1 (t_{1,i} | P_1, P_2, \gamma) + \frac{1}{n} \sum_{i=1}^{n} t_2 (x_{i} | P_1, P_2, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + O \left( \frac{\log n}{n} \right), t_1 \in B_N(P_1), x \in B_n(P_2) \right\}
\]
\[
\leq \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \mathbb{P}_2 \left\{ \frac{1}{n} \sum_{i=1}^{N} t_1 (t_{1,i} | P_1, P_2, \gamma) + \frac{1}{n} \sum_{i=1}^{n} t_2 (x_{i} | P_1, P_2, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + O \left( \frac{\log n}{n} \right) \right\}
\]
\[
= \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \mathbb{P}_2 \left\{ \sum_{i=1}^{N} \left\{ t_1 (t_{1,i} | P_1, P_2, \gamma) - \mathbb{E}_{P_1} [t_1 (X | P_1, P_2, \gamma)] \right\} + \sum_{i=1}^{n} \left\{ t_2 (x_{i} | P_1, P_2, \gamma) - \mathbb{E}_{P_2} [t_2 (X | P_1, P_2, \gamma)] \right\} \right\} \leq n \left( \lambda + \frac{r}{\sqrt{n}} - \text{GJS}(P_1, P_2, \gamma) \right) + O \left( \frac{\log n}{n} \right)
\]
where (A27) follows from Lemma 3 and the proof of (A28) is provided in Appendix F.

\[
\limsup_{n \to \infty} \beta_2(L^n | P_1, P_2)
\]
\[
\leq \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \left\{ G \left( \lambda + \frac{r}{\sqrt{n}} - \text{GJS}(P_1, P_2, \gamma) + O \left( \frac{\log n}{n} \right) \right) \sqrt{\frac{n}{V(P_1, P_2, \gamma)}} \right\}
\]
\[
\leq \limsup_{n \to \infty} \sum_{j=1}^{u} w(j) \left\{ G \left( \lambda + \frac{r}{\sqrt{n}} - \text{GJS}(P_1, P_2, \gamma) + O \left( \frac{\log n}{n} \right) \right) \sqrt{\frac{n}{V(P_1, P_2, \gamma)}} \right\}
\]
\[
\leq \sum_{j=1}^{u} w(j) \limsup_{n \to \infty} \left\{ G \left( \lambda + \frac{r}{\sqrt{n}} - \text{GJS}(P_1, P_2, \gamma) + O \left( \frac{\log n}{n} \right) \right) \sqrt{\frac{n}{V(P_1, P_2, \gamma)}} \right\},
\]
where \((A29)\) and the last inequality are derived from Lemma \(A1\) and Fatou’s lemma, respectively. Here,

\[
\limsup_{n \to \infty} G\left( \lambda + \frac{r}{\sqrt{n}} - \text{GJS}(P_1, P_j^2, \gamma) + O\left( \frac{\log n}{n} \right) \right) \sqrt{\frac{n}{V(P_1, P_j^2, \gamma)}} = \begin{cases} 
1 & \text{for } j \in U: \text{GJS}(P_1, P_j^2, \gamma) < \lambda \\
G\left( \frac{r}{\sqrt{V(P_1, P_j^2, \gamma)}} \right) & \text{for } j \in U: \text{GJS}(P_1, P_j^2, \gamma) = \lambda \\
0 & \text{for } j \in U: \text{GJS}(P_1, P_j^2, \gamma) > \lambda
\end{cases}.
\]

(A31)

Therefore,

\[
\limsup_{n \to \infty} \beta_2(\Lambda^n | P_1, P_2) \leq \epsilon.
\]

(A33)

**Appendix E.2. Converse Part**

For any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\) and for all pairs of distributions \((\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\), fix any test \(\phi^n\) satisfying that

\[
\beta_1(\phi^n | \tilde{P}_1, \tilde{P}_2) \leq \exp(-n\lambda - \sqrt{nr}) \quad \text{for } \lambda + \frac{r}{\sqrt{n}}.
\]

(A34)

\[
\limsup_{n \to \infty} \beta_2(\phi^n | P_1, P_2) \leq \epsilon.
\]

(A35)

We show

\[
\frac{r}{\sqrt{n}} + \sum_{\{j \in U : \text{GJS}(P_1, P_j^2, \gamma) < \lambda \}} w(j) + \sum_{\{j \in U : \text{GJS}(P_1, P_j^2, \gamma) = \lambda \}} \Phi(r) w(j) \leq \epsilon.
\]

(A36)

To prove (A36), we use Corollary 1 with replacing \(\lambda \) with \(\lambda + \frac{r}{\sqrt{n}}\). A lower bound on the type-II error probability of the test \(\phi^n\) for any pair of distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})\) can be evaluated as follows:
\[
\limsup_{n \to \infty} \beta_2(\phi^n | P_1, P_2) \\
\geq \limsup_{n \to \infty} \left( 1 - \frac{1}{n} \sum_{j=1}^n w(j)P_2^j \left\{ GJS(q_t, q_x, \gamma) + O \left( \frac{\log n}{n} \right) \right\} \right) + \rho(n) \geq \lambda + \frac{r}{\sqrt{n}}
\]

\[
\geq \limsup_{n \to \infty} \left( 1 - \frac{1}{n} \sum_{j=1}^n w(j)P_2^j \left\{ \frac{1}{n} \sum_{i=1}^n t_1(t_{1i}, P_1, P_2^j, \gamma) + \frac{1}{n} \sum_{i=1}^n t_2(x_i, P_1, P_2^j, \gamma) \right\} \right) + O \left( \frac{\log n}{n} \right) \geq \lambda + \frac{r}{\sqrt{n}}.
\]

\[
t_1 \in B_N(P_1), \quad x \in B_n(P_2^j) \}
\]

\[
\limsup_{n \to \infty} \sum_{j=1}^n w(j) \left\{ G \left( \left[ \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_2^j, \gamma) + O \left( \frac{\log n}{n} \right) \right] \sqrt{\frac{n}{V(P_1, P_2^j, \gamma)}} \right) \right\}
\]

\[
= \liminf_{n \to \infty} \sum_{j=1}^n w(j) G \left( \left[ \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_2^j, \gamma) + O \left( \frac{\log n}{n} \right) \right] \sqrt{\frac{n}{V(P_1, P_2^j, \gamma)}} \right).
\]

\[
\geq \liminf_{n \to \infty} \sum_{j=1}^n w(j) G \left( \left[ \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_2^j, \gamma) + O \left( \frac{\log n}{n} \right) \right] \sqrt{\frac{n}{V(P_1, P_2^j, \gamma)}} \right),
\]

where (A37) and (A38) follow from Lemma 3 and Lemma A1, respectively. The last inequality is due to Fatou’s lemma. Here,

\[
\liminf_{n \to \infty} G \left( \left[ \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_2^j, \gamma) + O \left( \frac{\log n}{n} \right) \right] \sqrt{\frac{n}{V(P_1, P_2^j, \gamma)}} \right)
\]

\[
= \begin{cases} 
1 & \text{for } j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) < \lambda \\
\frac{r}{\sqrt{V(P_1, P_2^j, \gamma)}} & \text{for } j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) = \lambda \\
0 & \text{for } j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) > \lambda 
\end{cases}
\]

Therefore,

\[
\limsup_{n \to \infty} \beta_2(\phi^n | P_1, P_2)
\]

\[
\geq \sum_{\{j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) < \lambda\}} w(j) + \sum_{\{j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) = \lambda\}} G \left( \frac{r}{\sqrt{V(P_1, P_2^j, \gamma)}} \right) w(j)
\]

\[
= \sum_{\{j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) < \lambda\}} w(j) + \sum_{\{j \in \mathcal{U} : GJS(P_1, P_2^j, \gamma) = \lambda\}} \Phi(r) w(j).
\]
In view of (A35), Equation (A41) indicates that

\[
r \leq \sup \left\{ r \left| \sum_{j \in \Omega : GJS(P_1, P_2, \gamma) < \lambda} w(j) + \sum_{j \in \Omega : GJS(P_1, P_2, \gamma) = \lambda} \Phi_j(r) w(j) \leq \epsilon \right. \right\}. \tag{A42}
\]

\[\square\]

**Appendix F**

**Proof of Equation (A28).** We shall show

\[
\frac{1}{n} \sum_{i=1}^{N} t_1(t_{1,i}|P_1, P_{2, i}^j, \gamma) + \frac{1}{n} \sum_{i=1}^{n} t_2(x_i|P_1, P_{2, i}^j, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + O \left( \frac{\log n}{n} \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{N} \left\{ t_1(t_{1,i}|P_1, P_{2, i}^j, \gamma) - \mathbb{E}_{P_1}[t_1(X|P_1, P_{2, i}^j, \gamma)] \right\} + \frac{1}{n} \sum_{i=1}^{n} \left\{ t_2(x_i|P_1, P_{2, i}^j, \gamma) - \mathbb{E}_{P_2}[t_2(X|P_1, P_{2, i}^j, \gamma)] \right\} \leq \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_{2, i}^j, \gamma) + O \left( \frac{\log n}{n} \right). \tag{A43}
\]

The left-hand side of (A43) can be rewritten as follows:

\[
\frac{1}{n} \sum_{i=1}^{N} t_1(t_{1,i}|P_1, P_{2, i}^j, \gamma) + \frac{1}{n} \sum_{i=1}^{n} t_2(x_i|P_1, P_{2, i}^j, \gamma) \leq \lambda + \frac{r}{\sqrt{n}} + O \left( \frac{\log n}{n} \right)
\]

\[
\iff \frac{1}{n} \sum_{i=1}^{N} t_1(t_{1,i}|P_1, P_{2, i}^j, \gamma) - \gamma \mathbb{E}_{P_1}[t_1(X|P_1, P_{2, i}^j, \gamma)] + \frac{1}{n} \sum_{i=1}^{n} t_2(x_i|P_1, P_{2, i}^j, \gamma) - \mathbb{E}_{P_2}[t_2(X|P_1, P_{2, i}^j, \gamma)] \leq \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_{2, i}^j, \gamma) + O \left( \frac{\log n}{n} \right). \tag{A44}
\]

\[
\iff \frac{1}{n} \sum_{i=1}^{N} t_1(t_{1,i}|P_1, P_{2, i}^j, \gamma) - \frac{1}{n} \sum_{i=1}^{N} \mathbb{E}_{P_1}[t_1(X|P_1, P_{2, i}^j, \gamma)] + \frac{1}{n} \sum_{i=1}^{n} t_2(x_i|P_1, P_{2, i}^j, \gamma) - \mathbb{E}_{P_2}[t_2(X|P_1, P_{2, i}^j, \gamma)] \leq \lambda + \frac{r}{\sqrt{n}} - GJS(P_1, P_{2, i}^j, \gamma) + O \left( \frac{\log n}{n} \right). \tag{A45}
\]

where (A44) follows from (47). Therefore, we obtain Equation (A28). \(\square\)

**References**

1. Merhav, N.; Ziv, J. A Bayesian approach for classification of Markov sources. *IEEE Trans. Inf. Theory* **1991**, *37*, 1067–1071. [CrossRef]
2. Saito, S.; Matsushima, T. Evaluation of error probability of classification based on the analysis of the Bayes code. In *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, 21–26 June 2020; pp. 21–26.
3. Gutman, M. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. Inf. Theory* **1989**, *35*, 401–408. [CrossRef] [PubMed]
4. Hsu, H.-W.; Wang, I.-H. On binary statistical classification from mismatched empirically observed statistics. In *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, 21–26 June 2020; pp. 3533–2538.
5. Zhou, L.; Tan, V.Y.F.; Motani, M. Second-order asymptotically optimal statistical classification. *Inf. Inference J. IMA* **2020**, *9*, 81–111.
6. Han, T.S.; Nomura, R. First- and second-order hypothesis testing for mixed memoryless sources. *Entropy* **2018**, *20*, 174. [CrossRef] [PubMed]
7. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
8. Polyanskiy, Y.; Poor, H.V.; Verdú, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
9. Yagi, H.; Han, T.S.; Nomura, R. First- and second-order coding theorems for mixed memoryless channels with general mixture. *IEEE Trans. Inf. Theory* **2016**, *62*, 4395–4412. [CrossRef]
10. Ziv, J. On classification with empirically observed statistics and universal data compression. *IEEE Trans. Inf. Theory* **1988**, *34*, 278–286. [CrossRef]
11. Kelly, B.G.; Wagner, A.B.; Tularak, T.; Viswanath, P. Classification of homogeneous data with large alphabets. *IEEE Trans. Inf. Theory* **2013**, *59*, 782–795. [CrossRef]
12. Unnikrishnan, J.; Huang, D. Weak convergence analysis of asymptotically optimal hypothesis tests. *IEEE Trans. Inf. Theory* **2016**, *62*, 4285–4299. [CrossRef]
13. He, H.; Zhou, L.; Tan, V.Y.F. Distributed detection with empirically observed statistics. *IEEE Trans. Inf. Theory* **2020**, *66*, 4349–4367. [CrossRef]
14. Saito, S.; Matsushima, T. Evaluation of error probability of classification based on the analysis of the Bayes code: extension and example. In *Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT)*, Melbourne, VIC, Australia, 12–20 July 2021; pp. 1445–1450.
15. Csiszár, I. The method of types. *IEEE Trans. Inf. Theory* **1998**, *44*, 2505–2523. [CrossRef]
16. Nielsen, F. On a variational definition for the Jensen-Shannon symmetrization of distances based on the information radius. *Entropy* **2021**, *23*, 464. [CrossRef] [PubMed]