Neyman-Pearson Test for Zero-Rate Multiterminal Hypothesis Testing

Shun Watanabe

Abstract

The problem of zero-rate multiterminal hypothesis testing is revisited from the perspective of information-spectrum approach and finite blocklength analysis. A Neyman-Pearson-like test is proposed and its non-asymptotic performance is clarified; for a short block length, it is numerically determined that the proposed test is superior to the previously reported Hoeffding-like test proposed by Han-Kobayashi. For a large deviation regime, it is shown that our proposed test achieves an optimal trade-off between the type I and type II exponents presented by Han-Kobayashi. Among the class of symmetric (type based) testing schemes, when the type I error probability is non-vanishing, the proposed test is optimal up to the second-order term of the type II error exponent; the latter term is characterized in terms of the variance of the projected relative entropy density. The information geometry method plays an important role in the analysis as well as the construction of the test.

I. INTRODUCTION

In the classic hypothesis testing problem, upon observing $Z^n$, a tester tries to distinguish whether the observation comes from the null hypothesis $P$ or the alternative hypothesis $Q$. It is widely known that the so-called Neyman-Pearson test [26] is the most powerful test in this regard and the trade-off between the type I error probability $\alpha_{NP}^n$ and the type II error probability $\beta_{NP}^n$ is characterized as

$$\alpha_{NP}^n = P\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda(Z_i) \leq \tau\right), \quad \beta_{NP}^n = Q\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda(Z_i) > \tau\right).$$

(1)

Here, $\Lambda(z) = t_{P||Q}(z) = \log \frac{P(z)}{Q(z)}$ is the log-likelihood ratio between the two distributions. This ratio is also known as the relative entropy density. An application of the law of large numbers to (1) implies that, for vanishing type I error probability, the asymptotically optimal exponent of the type II error probability is given by the relative

A part of this paper was presented in 2017 IEEE International Symposium on Information Theory.

S. Watanabe is with the Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Japan, E-mail:shunwata@cc.tuat.ac.jp.

1Technically speaking, for finite blocklength, non-randomized Neyman-Pearson test is optimal only for limited number of trade-off points, and randomization is required in general (eg. see [24]).

2Throughout the paper, we only consider discrete random variables taking values in finite sets. The notations $P(\cdot)$ and $Q(\cdot)$ in (1) mean that the probabilities of events are computed with respect to a sequence of i.i.d. (independent and identically distributed) random variables $Z^n$ that are distributed according to product distributions $P^n$ and $Q^n$, respectively. By a slight abuse of notation, we use the same notations $P$ and $Q$ to describe the probability mass functions.
entropy $D(P\|Q)$; more refined analyses on (1) give the tight bounds on more detailed asymptotics such as the large deviation regime or the second-order regime [9], [30].

Another important test, which we refer to as the Hoeffding test, entails comparing the type (empirical distribution) $t_{Z^n}$ of the observation with the null hypothesis [21]; the null hypothesis is accepted if the relative entropy between the type and $P$ is smaller than a prescribed threshold, and is rejected otherwise. The trade-off between the type I error probability $\alpha_{H_n}^I$ and the type II error probability $\beta_{H_n}^I$ of this test is characterized as

$$\alpha_{H_n}^I = P\left(D(t_{Z^n}\|P) \geq r\right), \quad \beta_{H_n}^I = Q\left(D(t_{Z^n}\|P) < r\right).$$

(2)

The advantage of this test is that it can be conducted without knowledge of $Q$, i.e., it is partially universal (see [12]). Although the Hoeffding test delivers optimal performance asymptotically [11, Theorem III.2], the trade-off between the errors is worse than that of the Neyman-Pearson test for a finite block length, in particular for a short block length.

In [8], Berger introduced a new framework of multiterminal statistical decision problems under communication constraint. Inspired by his work, many researchers studied various problems of this kind [11, 16, 18, 6, 2, 18, 4] (see [17] for a thorough review). One important special case of these problems is the zero-rate multiterminal hypothesis testing problem, which is the main topic of this paper. This problem involves separate processing of the correlated observations $X^n$ and $Y^n$ by two encoders, after which messages are sent to a centralized decoder at zero-rate. Then, the decoder tries to distinguish whether the observations originate from the null hypothesis $P_{XY}$ or the alternative hypothesis $Q_{XY}$. It was shown in [16], [29] that, for a type I error with vanishing probability, the asymptotically optimal exponent of the type II error probability is given by the projected relative entropy defined by

$$E(P_{XY}\|Q_{XY}) = \min_{\tilde{P}_{XY} : \tilde{P}_{X} = P_X, \tilde{P}_Y = P_Y} D(\tilde{P}_{XY}\|Q_{XY}).$$

(3)

In [6], Amari-Han studied this problem from a differential geometrical viewpoint, and provided a geometrical interpretation of (3) by using the information geometry approach [7], [5]. In fact, the term, projected relative entropy, should be clear from the observation in [6] (see Section III). Furthermore, to study the large deviation regime of the zero-rate multiterminal hypothesis testing problem, Han-Kobayashi introduced a Hoeffding-like testing scheme for this problem [18]; the trade-off in the error of their testing scheme is characterized as

$$\alpha_{H_l}^I = P\left(E(t_{X^nY^n}\|P_{XY}) \geq r\right), \quad \beta_{H_l}^I = Q\left(E(t_{X^nY^n}\|P_{XY}) < r\right),$$

(4)

where $t_{X^nY^n}$ is the joint type of $(X^n, Y^n)$. It was shown in [17] that the bound in (4) is asymptotically tight in a large deviation regime.

3More precisely, an achievability testing scheme was proposed in [16], and the converse for the so-called one-bit compression case was shown under some regularity condition. Later, the converse for the zero-rate compression case was shown in [29] under some regularity condition. Recently, the regularity condition of [29] was relaxed in [13]. Further, it is worth mentioning that the answer to the multiterminal hypothesis testing with a comparator is given by the same quantity, [9], [27].
Thus far, we have reviewed the background on the zero-rate multiterminal hypothesis testing problem. The main aim of this paper is to revisit this problem from the perspective of modern approaches developed in the past two decades, which are reviewed next.  

In their landmark paper [19], Han-Verdú proposed the information-spectrum approach. Among other aspects, a key feature of this approach is that the performance of a coding problem is characterized by the probabilistic behavior of the information density that is inherent to that coding problem. For instance, in the case of hypothesis testing, the relative entropy density $\mathcal{I}_{P\|Q}$ can be regarded as the information density of this problem. The same philosophy was inherited by another recently popularized area of research, the finite block length and the second-order analyses, pioneered by Hayashi [20] and Polyanskiy-Poor-Verdú [28]. In the second-order analyses, instead of the law of the large number, the central limit theorem is applied to analyze the probabilistic behavior of the information density up to $\sqrt{n}$ order.

Although information densities can be naturally introduced for some problems, it is non-trivial to identify the correct quantity in general. For instance, Kostina-Verdú introduced the $D$-tilted information density for the lossy source coding problem in a judicious manner, and characterized the second-order coding rate in terms of the variance of this information density [23] (see also [22] for an alternative approach proposed by Ingber-Kochman). The same direction of research was extended to the Gray-Wyner network, one of the most basic multiterminal problems, by the author in [32] (see also [34]).

As mentioned above, the classic hypothesis-testing problem comprises two important tests: the Neyman-Pearson and Hoeffding tests. The test proposed by Han-Kobayashi [18] can be regarded as a Hoeffding test for the zero-rate multiterminal hypothesis testing. Thus, it is tempting, both theoretically and practically, to have a testing scheme that is reminiscent of the Neyman-Pearson test. In this paper, we propose such a testing scheme. In fact, the trade-off between the type I and type II error probabilities by our proposed test has the following form:

$$
\alpha_n^{\text{HP1}} = P\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_{\lambda}(X_i, Y_i) \leq \tau\right), \quad \beta_n^{\text{HP1}} = Q\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_{\lambda}(X_i, Y_i) > \tau\right)
$$

Here, $\Lambda_{\lambda}(x, y)$ is a proxy of the log-likelihood ratio parametrized by $\lambda \in [-E(Q_{XY}\|P_{XY}), E(P_{XY}\|Q_{XY})]$; as is subsequently shown, identification of $\Lambda_{\lambda}(x, y)$ is non-trivial, which is one of technical contributions of this paper. In contrast to the Neyman-Pearson test in the classic hypothesis testing, the proxy of the log-likelihood is parametrized by $\lambda$. As we will see later in the paper, adjustment of $\lambda$ depending on threshold $\tau$ is very important. For instance, the optimal choice turns out to be $\lambda = \tau$ in the large deviation regime. An extreme case $\lambda = E(P_{XY}\|Q_{XY})$ of this proxy of the log-likelihood ratio, which we term the projected relative entropy density, is $J_{P\|Q}(x, y) = \log \frac{P_{XY}(x, y)}{Q_{XY}(x, y)}$ for the optimizer $P_{XY}^*$ of $E(P_{XY}\|Q_{XY})$ in (5). In fact, it will be clarified that the expectation of $J_{P\|Q}(x, y)$ over $P_{XY}$ is given by $E(P_{XY}\|Q_{XY})$.

\[8\]

At the time when the multiterminal hypothesis testing was actively studied in the late 80s to early 90s, it seems that the method of type [12] was the most popular tool for analysis.

Another key feature is its generality, i.e., coding theorems are proved without any assumptions such as stationarity and/or ergodicity.

As a rule of thumb, $\lambda = \tau$ provides the best error trade-off even in the finite blocklength regime.
Although it is not clear whether our proposed testing scheme is the most powerful, for a rather short block length, we numerically determine that our proposed testing scheme has better error trade-off than that of the previously known test of Han-Kobayashi. We also show that, for a large deviation regime, our proposed test achieves optimal trade-off between the type I and type II exponents shown by Han-Kobayashi. Furthermore, among the class of symmetric (type based) testing schemes, we derive the optimal second-order rate of the type II exponent, which can be achieved by our proposed test. Ultimately, it emerges that the optimal second-order rate is characterized by the variance of the projected relative entropy density.

Here, we would like to mention the similarity and dissimilarity of the Neyman-Pearson test in classic hypothesis testing and our Neyman-Pearson-like test in the multiterminal hypothesis testing from a geometrical viewpoint. In terms of the classic hypothesis-testing problem, the Neyman-Pearson test is known to correspond to bisecting the probability simplex by a mixture family generated by the log-likelihood ratio $\Lambda(z)$, which is orthogonal to the e-geodesic connecting the null hypothesis $P$ and the alternative hypothesis $Q$ (e.g., see [25]). On the other hand, our Neyman-Pearson-like test of the multiterminal hypothesis testing bisects the probability simplex by a mixture family generated by the proxy $\Lambda_\lambda(x, y)$ of the log-likelihood ratio. In contrast to the Neyman-Pearson test in classic hypothesis testing, our Neyman-Pearson-like test has the freedom to adjust the direction of bisection by parameter $\lambda$. Interestingly, this adjustment of direction is crucial to achieve the optimal trade-off between the type I and type II error exponents in a large deviation regime.

In addition to the above mentioned geometrical motivation, there is also a practical motivation to introduce the Neyman-Pearson-like test. Even for multiterminal hypothesis testing, we can conduct the Neyman-Pearson test by computing the log-likelihood ratio for a given observation. However, computing such a log-likelihood ratio is intractable as the blocklength become larger. On the other hand, our Neyman-Pearson-like scheme only requires computing the empirical average of the proxy of the log-likelihood ratio, and it is easier to implement (see Remark 5).

The remainder of the paper is organized as follows: In Section II we introduce our notation, and recall the problem formulation of the multiterminal hypothesis testing. We also review the previously known testing scheme of Han-Kobayashi. In Section III we review basic results on information geometry as well as the result of Amari-Han [6]. In Section IV we introduce our novel testing scheme. In Section V we consider a binary example, and compare the two testing schemes. In Section VI the large deviation performance of our proposed test is discussed. In Section VII we derive the second-order exponent. We conclude the paper with some discussions in Section VIII.

II. PROBLEM FORMULATION

A. Notation

Random variables (e.g., $X$) and their realizations (e.g., $x$) are presented in upper and lower case format, respectively. All random variables take values in some finite alphabets, which are denoted in calligraphic font (e.g., $X$). The cardinality of $X$ is denoted as $|X|$. Let the random vector $X^n = (X_1, \ldots, X_n)$ and similarly for its realization $x = (x_1, \ldots, x_n)$. For information theoretic quantities, we follow the same notations as [12]; e.g., the
entropy and the relative entropy are denoted by \(H(X)\) and \(D(P\|Q)\), respectively. For two distributions \(P\) and \(P'\), the variational distance is denoted by \(||P-P'||\). For a sequence \(x\), its type (empirical distribution) is denoted by \(t_x\); similarly the joint type of \((x,y)\) is denoted by \(t_{xy}\). The set of all positive probability distributions on \(X\) is denoted by \(\mathcal{P}(X)\), the set of all probability distributions (not necessarily positive) is denoted by \(\overline{\mathcal{P}}(X)\), and the set of all types (not necessarily positive) on \(X\) is denoted by \(\mathcal{P}_n(X)\). Similar notations are used for the sets of joint distributions and joint types. In addition, the notations \(E[\cdot]\) and \(\text{Var}[\cdot]\) mean computing the expectation and the variance of the random variables in the bracket, respectively. \(\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du\) is the cumulative distribution function of the standard normal distribution; its inverse is denoted by \(\Phi^{-1}(\varepsilon)\) for \(0 < \varepsilon < 1\). Throughout the paper, the base of \(\log\) and \(\exp\) are \(e\).

B. Multiterminal Hypothesis Testing

In this section, we introduce the problem setting and review some basic results. We consider the statistical problem of testing the null hypothesis \(H_0 : P_{XY}\) on \(X \times Y\) versus the alternative hypothesis \(H_1 : Q_{XY}\) on the same alphabet. We assume \(P_{XY}\) and \(Q_{XY}\) are positive, i.e., they have full support throughout the paper.\(^7\) The i.i.d. random variables \((X^n, Y^n)\) distributed according to either \(P^n_{XY}\) or \(Q^n_{XY}\) are observed separately by two terminals, and they are encoded by two encoders

\[
\begin{align*}
    f_1^{(n)} : X^n &\to \mathcal{M}_1^{(n)}, \\
    f_2^{(n)} : Y^n &\to \mathcal{M}_2^{(n)},
\end{align*}
\]

respectively. Then, the decoder

\[
g^{(n)} : \mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)} \to \{H_0, H_1\}
\]

decides whether to accept the null hypothesis. When the block length \(n\) is obvious from the context, we omit the superscript \(n\). For a given testing scheme \(T_n = (f_1, f_2, g)\), the type I error probability is defined by

\[
\alpha[T_n] := P\left( g(f_1(X^n), f_2(Y^n)) = H_1 \right)
\]

and the type II error probability is defined by

\[
\beta[T_n] := Q\left( g(f_1(X^n), f_2(Y^n)) = H_0 \right).
\]

In the remaining part of the paper, \(P(\cdot)\) (or \(Q(\cdot)\)) means \((X^n, Y^n)\) is distributed according to \(P^n_{XY}\) (or \(Q^n_{XY}\)).

A sequence of the testing scheme \(\{T_n\}_{n=1}^\infty\) is said to be zero-rate if

\[
\lim_{n \to \infty} \frac{1}{n} \log |\mathcal{M}_i^{(n)}| = 0, \; i = 1, 2.
\]

Furthermore, \(\{T_n\}_{n=1}^\infty\) is said to be a symmetric testing scheme if the encoders \(f_1^{(n)}\) and \(f_2^{(n)}\) only depend on the marginal types of \(x\) and \(y\), respectively. The class of zero-rate schemes was introduced in \cite{16} and the class\(^7\) The results on multiterminal hypothesis testing \cite{29}, \cite{17} are sensitive to the supports of the alternative hypothesis. Even though the full support assumption may be slightly relaxed (eg. see \cite{31}), we only consider the full support case in this paper.
of symmetric schemes was introduced in [6]. Note that a symmetric scheme is automatically zero-rate since the number of marginal types is polynomial in $n$.

Let

$$E_0(P_{XY} || Q_{XY}) := \sup \left\{ \liminf_{n \to \infty} -\frac{1}{n} \log \beta[T_n] : \{T_n\}_{n=1}^{\infty} \text{ is zero-rate, } \lim_{n \to \infty} \alpha[T_n] = 0 \right\}$$

(12)

and

$$E_s(P_{XY} || Q_{XY}) := \sup \left\{ \liminf_{n \to \infty} -\frac{1}{n} \log \beta[T_n] : \{T_n\}_{n=1}^{\infty} \text{ is symmetric, } \lim_{n \to \infty} \alpha[T_n] = 0 \right\}$$

(13)

be the optimal exponent of the type II error probability in each class of schemes. By definition, $E_0(P_{XY} || Q_{XY}) \geq E_s(P_{XY} || Q_{XY})$. These quantities can be characterized as follows.

**Proposition 1 ([16], [29])** It holds that

$$E_0(P_{XY} || Q_{XY}) = E_s(P_{XY} || Q_{XY}) = E(P_{XY} || Q_{XY}),$$

(14)

where

$$E(P_{XY} || Q_{XY}) := \min \left\{ D(\tilde{P}_{XY} || Q_{XY}) : \tilde{P}_{XY} \in \mathcal{P}(X \times Y), \tilde{P}_X = P_X, \tilde{P}_Y = P_Y \right\}.$$  

(15)

**C. Han-Kobayashi Testing Scheme**

In [18], Han-Kobayashi studied a large deviation regime of multiterminal hypothesis testing. For $0 \leq r \leq E(Q_{XY} || P_{XY})$, let

$$F(r) := \sup \left\{ \liminf_{n \to \infty} -\frac{1}{n} \log \beta[T_n] : \{T_n\}_{n=1}^{\infty} \text{ is zero-rate, } \liminf_{n \to \infty} -\frac{1}{n} \log \alpha[T_n] \geq r \right\}.$$  

(16)

To derive a lower bound on $F(r)$, Han-Kobayashi proposed the following Hoeffding-like testing scheme:

By definition, note that

$$E(P_{XY} || P_{XY}) = E(P_{\bar{X}} \times P_{\bar{Y}} || P_{XY})$$

(17)

holds for any joint distribution $P_{XY}$, where $P_{\bar{X}} \times P_{\bar{Y}}$ are the product distribution of the marginals $P_X, P_Y$ of $P_{XY}$. Upon observing $x$ and $y$, the encoders send their types. Then, upon receiving a pair of marginal types $(t_x, t_y)$, the decoder computes $E(t_x \times t_y || P_{XY})$; if the value is smaller than a prescribed threshold $r$, then it outputs $H_0$; otherwise, it outputs $H_1$. By (17), $g(f_1(x), f_2(y)) = H_0$ if and only if

$$E(t_{xy} || P_{XY}) < r.$$  

(18)

In fact, the threshold $r$ controls the convergence speed of the type I error probability, i.e., in the large deviation regime, the type I error probability behaves as $\exp\{-nr\}$. Apparently, this scheme is a symmetric scheme. The performance of this scheme is summarized in the following proposition.

---

8The exponents defined in (12) and (13) are the so-called weak converse, i.e., we require that the type I error probability converges to 0. In fact, it is known that the strong converse holds for this problem [29].

9The testing scheme proposed in [18] is slightly different, but it is essentially the same as the scheme reviewed in this section.
Proposition 2 For a given \( r > 0 \), the above mentioned Hoeffding like scheme \( T_n^{\text{HL}} \) has the following error trade-off:

\[
\alpha[T_n^{\text{HL}}] = P\left(E(t_{X^nY^n} \parallel P_{XY}) \geq r\right), \tag{19}
\]

\[
\beta[T_n^{\text{HL}}] = Q\left(E(t_{X^nY^n} \parallel P_{XY}) < r\right). \tag{20}
\]

It was shown in [18] that the above testing scheme satisfies

\[
\lim_{n \to \infty} -\frac{1}{n} \log \alpha[T_n^{\text{HL}}] = r \tag{21}
\]

and

\[
\lim_{n \to \infty} -\frac{1}{n} \log \beta[T_n^{\text{HL}}] = \min \left\{ D(Q_{XY} \parallel P_{XY}) : E(P_{XY} \parallel P_{XY}) \leq r \right\} \tag{22}
\]

\[
= \min \left\{ E(P_{XY} \parallel Q_{XY}) : D(P_{XY} \parallel P_{XY}) \leq r \right\}, \tag{23}
\]

which is optimal among the class of all zero-rate testing schemes [17]; we summarize these results in the following proposition.

Proposition 3 ([18], [17]) For \( 0 \leq r \leq E(Q_{XY} \parallel P_{XY}) \), it holds that

\[
F(r) = \min \left\{ D(Q_{XY} \parallel P_{XY}) : E(Q_{XY} \parallel P_{XY}) \leq r \right\} \tag{24}
\]

\[
= \min \left\{ E(P_{XY} \parallel Q_{XY}) : D(P_{XY} \parallel P_{XY}) \leq r \right\}. \tag{25}
\]

III. Preliminaries of Information Geometry

In this section, we review some results on information geometry that are needed in later sections. Interested readers are referred to [7], [13] for a thorough review on information geometry.

A. Properties of Projection

In this section, we review some properties of projection with respect to the relative entropy, which is sometimes known as I-projection. Let \( C \subseteq \mathcal{P}(Z) \) be a (nonempty) closed convex set. For a given \( Q \in \mathcal{P}(Z) \), let us consider the following optimization problem:

\[
\min_{\tilde{P} \in C} D(\tilde{P} \parallel Q). \tag{26}
\]

The optimizer of (26) satisfies the following extremal condition [10] (see also [13, Section 3]).

Theorem 1 ([10]) The optimizer \( P^* \) of (26) is unique; furthermore, for every \( \tilde{P} \in C \), the optimizer \( P^* \) satisfies \( \text{supp}(\tilde{P}) \subseteq \text{supp}(P^*) \) and

\[
D(\tilde{P} \parallel Q) \geq D(\tilde{P} \parallel P^*) + D(P^* \parallel Q). \tag{27}
\]

In fact, only the expression (24) for \( F(r) \) was derived in [18], [17]; however, the expression (25) can be also described by the expression (24). Later in Lemma 2 we will verify that the two expressions coincide for \( 0 < r < E(Q_{XY} \parallel P_{XY}) \); for \( r = 0 \) and \( r = E(Q_{XY} \parallel P_{XY}) \), it is not difficult to see that the two expressions coincide and are given by \( E(P_{XY} \parallel Q_{XY}) \) and 0, respectively.
Note that (27) is equivalent to
\[ \sum_z (\tilde{P}(z) - P^*(z)) \log \frac{P^*(z)}{Q(z)} \geq 0. \] 
(28)

The set
\[ S_*(P^*) := \left\{ \tilde{P} : \sum_z (\tilde{P}(z) - P^*(z)) \log \frac{P^*(z)}{Q(z)} = 0 \right\} \] 
(29)
is the tangent plane of \( C \) at \( P^* \) in the sense of (28). Moreover, we have
\[ C \subseteq S_{\geq}(P^*) := \left\{ \tilde{P} : \sum_z (\tilde{P}(z) - P^*(z)) \log \frac{P^*(z)}{Q(z)} \geq 0 \right\} \] 
(30)
and
\[ \min_{\tilde{P} \in S_{\geq}(P^*)} D(\tilde{P}||Q) = D(P^*||Q). \] 
(31)

For given functions \( f_1, \ldots, f_k \) from \( Z \) to \( \mathbb{R} \) and constants \( c_1, \ldots, c_k \in \mathbb{R} \), the set
\[ M := \left\{ \tilde{P} \in \mathcal{P}(Z) : \sum_z (\tilde{P}(z)) f_i(z) = c_i, \forall 1 \leq i \leq k \right\} \] 
(32)
is known as a mixture family. We make the assumption that \( M \) is not empty, and that it contains at least one element \( \tilde{P} \) having full support. On the other hand, the set \( \mathcal{E} \) of all distributions of the form
\[ \tilde{P}(z) = Q(z) \exp \left[ \sum_{i=1}^{k} \theta f_i(z) - \psi(\theta) \right], \quad \theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k \] 
(33)
is termed the exponential family generated by \( Q \) and \( f_1, \ldots, f_k \); the normalization constant \( \psi(\theta) \) is usually known as a potential function. When \( C = M \), the optimizer of (26) satisfies the following Pythagorean identity, and the optimizer is included in the exponential family \( \mathcal{E} \).

**Theorem 2** ([10]) When \( C = M \) is a mixture family, the optimizer \( P^* \) of (26) satisfies
\[ D(\tilde{P}||Q) = D(\tilde{P}||P^*) + D(P^*||Q) \] 
(34)
for every \( \tilde{P} \in M \), and \( P^* \in \mathcal{E} \cap M \).

**B. Geometry of \( \mathcal{P}(X \times Y) \)**

In this section, we review the results in [6] (see also [3, 7]). We first introduce a coordinate system on the set of all positive joint distributions, \( \mathcal{P}(X \times Y) \). Note that \( \mathcal{P}(X \times Y) \) is a \((|X||Y| - 1)\)-dimensional manifold. When we consider a multiterminal problem, it is convenient to consider the following parametrization specified by \( (d_x d_y + d_x + d_y) \) parameters, where \( d_x = |X| - 1 \) and \( d_y = |Y| - 1 \). Henceforth, we identify the alphabets as \( X = \{0, 1, \ldots, d_x\} \) and \( Y = \{0, 1, \ldots, d_y\} \). First, a natural parameter is introduced as
\[ P_{X,Y,\theta}(x, y) := \exp \left[ \sum_{i=1}^{d_x} \theta_{i}^x \delta_i(x) + \sum_{j=1}^{d_y} \theta_{j}^y \delta_j(y) + \sum_{i=1}^{d_x} \sum_{j=1}^{d_y} \theta_{ij}^x \delta_{ij}(x, y) - \psi(\theta) \right], \] 
(35)

\[ \text{Sometimes, it is also referred to as a linear family.} \]
where \( \delta_i(x) = 1[x = i] \) is the Kronecker delta (\( \delta_j(y) \) and \( \delta_i(x, y) \) are defined similarly),

\[
\theta^x_i = \log \frac{P_{XY, \theta}(i, 0)}{P_{XY, \theta}(0, 0)},
\]

\[
\theta^y_j = \log \frac{P_{XY, \theta}(0, j)}{P_{XY, \theta}(0, 0)},
\]

\[
\theta^\gamma_{ij} = \log \frac{P_{XY, \theta}(i, j)P_{XY, \theta}(0, 0)}{P_{XY, \theta}(i, 0)P_{XY, \theta}(0, j)}
\]

for \( 1 \leq i \leq d_x \) and \( 1 \leq j \leq d_y \), and \( \psi(\theta) \) is the normalization constant (potential function) given by

\[
\psi(\theta) = -\log P_{XY, \theta}(0, 0)
\]

\[
= \log \left( 1 + \sum_{i=1}^{d_x} e^{\theta^x_i} + \sum_{j=1}^{d_y} e^{\theta^y_j} + \sum_{i=1}^{d_x} \sum_{j=1}^{d_y} e^{\theta^x_i+\theta^\gamma_{ij}} \right).
\]

In this coordinate system, \( \{\theta^\gamma_{ij}\} \) describe the correlation between \( X \) and \( Y \). In fact, we can verify that \( P_{XY, \theta} = P_{X, \theta} \times P_{Y, \theta} \) if and only if \( \theta^\gamma_{ij} = 0 \) for every \( 1 \leq i \leq d_x \) and \( 1 \leq j \leq d_y \).

Next, the expectation parameter is introduced as

\[
P_{XY, \eta}(x, y) := \sum_{i=1}^{d_x} \eta^x_i (\delta_i(x) - \delta_0(x)) \delta_0(y) + \sum_{j=1}^{d_y} \eta^y_j \delta_0(x) (\delta_j(y) - \delta_0(y))
\]

\[
+ \sum_{i=1}^{d_x} \sum_{j=1}^{d_y} \eta^\gamma_{ij} (\delta_{ij}(x, y) - \delta_i(x) \delta_0(y) - \delta_0(x) \delta_j(y) + \delta_0(x) \delta_0(y)) + \delta_0(x) \delta_0(y),
\]

where

\[
\eta^x_i = P_{X, \eta}(i),
\]

\[
\eta^y_j = P_{Y, \eta}(j),
\]

\[
\eta^\gamma_{ij} = P_{XY, \eta}(i, j)
\]

for \( 1 \leq i \leq d_x \) and \( 1 \leq j \leq d_y \). The parameters \( \{\eta^x_i\} \) and \( \{\eta^y_j\} \) are called expectation parameters since \( \eta^x_i = \mathbb{E}[\delta_i(X)] \), \( \eta^y_j = \mathbb{E}[\delta_j(Y)] \), and \( \eta^\gamma_{ij} = \mathbb{E}[\delta_{ij}(X, Y)] \) for \( (X, Y) \sim P_{XY, \eta} \). Apparently in this coordinate system, \( \{\eta^x_i\} \) and \( \{\eta^y_j\} \) describe the marginal distributions, respectively.

By taking the derivative of (40), we have

\[
\frac{\partial \psi(\theta)}{\partial \theta^x_i} = \exp \left[ \theta^x_i - \psi(\theta) \right] + \sum_{j=1}^{d_y} \exp \left[ \theta^y_j + \theta^\gamma_{ij} - \psi(\theta) \right]
\]

\[
= \sum_{j=0}^{d_y} P_{XY, \theta}(i, j)
\]

\[
= P_{X, \theta}(i)
\]

for \( 1 \leq i \leq d_x \). Similarly, we have

\[
\frac{\partial \psi(\theta)}{\partial \theta^y_j} = P_{Y, \theta}(j)
\]
for $1 \leq j \leq d_y$, and

$$\frac{\partial \psi(\theta)}{\partial \theta^y_{ij}} = \exp \left[ \theta^x_i + \theta^y_j + \theta^y_{ij} - \psi(\theta) \right]$$

$$= P_{XY,\theta}(i, j)$$

for $1 \leq i \leq d_x$ and $1 \leq j \leq d_y$. Thus, the two coordinate systems $\theta$ and $\eta$ are related by

$$\eta^x_i(\theta) = \frac{\partial \psi(\theta)}{\partial \theta^x_i},$$

$$\eta^y_j(\theta) = \frac{\partial \psi(\theta)}{\partial \theta^y_j},$$

$$\eta^y_{ij}(\theta) = \frac{\partial \psi(\theta)}{\partial \theta^y_{ij}}.$$  (54)

Hereafter, we use notations such as $\theta^y = (\theta^y_{ij} : 1 \leq i \leq d_x, 1 \leq j \leq d_y)$ and $\eta^y = (\eta^y_{ij} : 1 \leq j \leq d_y)$ for brevity.

Let $\theta(P)$ and $\eta(P)$ be the natural and expectation parameters that correspond to $P_{XY}$, and let $\theta(Q)$ and $\eta(Q)$ be the natural and expectation parameters that correspond to $Q_{XY}$. Let

$$E(\theta^y(Q)) := \{ P_{XY,\theta} : \theta^y = \theta^y(Q) \}$$

be the exponential family containing $Q_{XY}$, and let

$$\mathcal{M}(\eta^x(P), \eta^y(P)) := \{ P_{XY,\theta} : \eta^x(\theta) = \eta^x(P), \eta^y(\theta) = \eta^y(P) \}$$

be the mixture family containing $P_{XY}$. Similarly, we define $E(\theta^y(P))$ and $\mathcal{M}(\eta^x(Q), \eta^y(Q))$ by replacing the roles of $P_{XY}$ and $Q_{XY}$. It was shown in [6] that the optimization problem in $E(P_{XY}\|Q_{XY})$ is achieved by the intersection of $E(\theta^y(Q))$ and $\mathcal{M}(\eta^x(P), \eta^y(P))$, and the interpretation of its Pythagorean theorem was given (see Fig. 1).  

**Theorem 3** ([6]) The optimizer $P^*_{XY}$ of $E(P_{XY}\|Q_{XY})$ satisfies $P^*_{XY} \in E(\theta^y(Q)) \cap \mathcal{M}(\eta^x(P), \eta^y(P))$ and

$$D(P_{XY}\|Q_{XY}) = D(P_{XY}\|P^*_{XY}) + D(P^*_{XY}\|Q_{XY}).$$

(57)

Similarly, the optimizer $Q^*_{XY}$ of $E(Q_{XY}\|P_{XY})$ satisfies $Q^*_{XY} \in E(\theta^y(P)) \cap \mathcal{M}(\eta^x(Q), \eta^y(Q))$ and

$$D(Q_{XY}\|P_{XY}) = D(Q_{XY}\|Q^*_{XY}) + D(Q^*_{XY}\|P_{XY}).$$

(58)

**Theorem 3** implies that $E(P_{XY}\|Q_{XY})$ is a projected component of the relative entropy $D(P_{XY}\|Q_{XY})$; thus, we term it the projected relative entropy.

For later use, we introduce a simple implication of Theorem 3.

**Corollary 1** For any $\tilde{P}_{XY} \in E(\theta^y(P))$ and $\tilde{Q}_{XY} \in E(\theta^y(Q))$ such that $\tilde{P}_X = \tilde{Q}_X$ and $\tilde{P}_Y = \tilde{Q}_Y$, it holds that

$$\sum_{x,y} \tilde{P}_{XY}(x,y) \log \frac{\tilde{Q}_{XY}(x,y)}{\tilde{Q}_{XY}(x,y)} = D(\tilde{Q}_{XY}\|Q_{XY}).$$

(59)

In fact, Theorem 3 is essentially a special case of Theorem 2. We reviewed both of these claims for later convenience.
Fig. 1: A description of Theorem 3. The mixture family $M(\eta^x(P), \eta^y(P))$ containing $P_{XY}$ and the exponential family $E(\theta^y(Q))$ containing $Q_{XY}$ intersect at $P^*_{XY}$. On the other hand, the mixture family $M(\eta^x(Q), \eta^y(Q))$ containing $Q_{XY}$ and the exponential family $E(\theta^y(P))$ containing $P_{XY}$ intersect at $Q^*_{XY}$. Those intersections are orthogonal in the sense of (57) and (58), respectively.

**Proof:** By applying Theorem 3 for $(\tilde{P}_{XY}, \tilde{Q}_{XY})$ in the place of $(P_{XY}, P^*_{XY})$, we have

$$D(\hat{Q}_{XY} \parallel Q_{XY}) = D(\hat{P}_{XY} \parallel Q_{XY}) - D(\hat{P}_{XY} \parallel \hat{Q}_{XY})$$

$$= \sum_{x,y} \hat{P}_{XY}(x,y) \left[ \log \frac{\hat{P}_{XY}(x,y)}{Q_{XY}(x,y)} - \log \frac{\hat{P}_{XY}(x,y)}{\hat{Q}_{XY}(x,y)} \right]$$

$$= \sum_{x,y} \hat{P}_{XY}(x,y) \log \frac{\hat{Q}_{XY}(x,y)}{Q_{XY}(x,y)}. \quad (62)$$

**IV. NEYMAN-PEARSON-LIKE TESTING SCHEME**

In this section, we propose a new testing scheme and evaluate its non-asymptotic performance.

We first investigate the properties of the exponent function $F(r)$. For that purpose, the following property of $E(P_{XY} \parallel Q_{XY})$ is useful.

**Lemma 1** $E(P_{XY} \parallel Q_{XY})$ is a convex function with respect to $P_{XY}$.

**Proof:** For given $\tilde{P}_{XY}$ and $\hat{P}_{XY}$ and $0 < p < 1$, we have

$$pE(\tilde{P}_{XY} \parallel Q_{XY}) + (1-p)E(\hat{P}_{XY} \parallel Q_{XY}) = pD(\tilde{P}_{XY} \parallel Q_{XY}) + (1-p)D(\hat{P}_{XY} \parallel Q_{XY}) \quad (63)$$

$$\geq D(p\tilde{P}_{XY} + (1-p)\hat{P}_{XY} \parallel Q_{XY}) \quad (64)$$

$$\geq E(p\tilde{P}_{XY} + (1-p)\hat{P}_{XY} \parallel Q_{XY}). \quad (65)$$
where $P_{XY}^*$ and $\tilde{P}_{XY}$ are optimizers of $E(P_{XY}||Q_{XY})$ and $E(\tilde{P}_{XY}||Q_{XY})$, respectively, and the last inequality holds since the marginals of $pP_{XY} + (1-p)\tilde{P}_{XY}$ and $pP_{XY} + (1-p)\tilde{P}_{XY}$ coincide.

Lemma 1 implies the set
\[ C := \{ \tilde{P}_{XY} : E(\tilde{P}_{XY}||P_{XY}) \leq r \}. \]
(66)
is a convex set. We can also find that $F(0) = E(P_{XY}||Q_{XY})$ and $F(E(Q_{XY}||P_{XY})) = 0$. $F(r)$ is also convex. For $0 < r < E(Q_{XY}||P_{XY})$, we have $F(r) > 0$. Thus, $F(r)$ is a monotonically decreasing function for $0 \leq r \leq E(Q_{XY}||P_{XY})$, and the minimization is attained at the boundary, i.e., $\tilde{P}_{XY}$ satisfying $E(\tilde{P}_{XY}||P_{XY}) = r$. Let
\begin{equation}
\lambda(r) := -r + F(r).
\end{equation}
(67)Since $F(r)$ is monotonically decreasing, $\lambda(r)$ is a one-to-one mapping for $0 \leq r \leq E(Q_{XY}||P_{XY})$, and we find that $\lambda(0) = E(P_{XY}||Q_{XY})$ and $\lambda(E(Q_{XY}||P_{XY})) = -E(Q_{XY}||P_{XY})$.

The following lemma provide some properties of the optimizers in (24) and (25).

**Lemma 2** For $0 < r < E(Q_{XY}||P_{XY})$, the optimizer $Q_X^*$ of (24) and the optimizer $P_{XY}^*$ of (25) are unique, and those optimizers satisfy $(Q_{XY}^*, P_{XY}^*) \in \mathcal{E}(\theta^\gamma(Q)) \times \mathcal{E}(\theta^\gamma(P))$ and have the same marginals, i.e., $Q_X^* = P_X^*$ and $Q_Y^* = P_Y^*$.

**Proof:** For simplicity of notation, we denote the values of (24) and (25) by $v_1$ and $v_2$, respectively. Since the set $C$ defined in (66) is a convex set, Theorem 1 implies that there exists a unique optimizer $Q_{XY}^*$ of (24) \footnote{Since $C$ contains $P_{XY}$, which has full support, Theorem 1 implies that $Q_{XY}^*$ has full support.} Note that $Q_{XY}^*$ must be included in $\mathcal{E}(\theta^\gamma(Q))$; otherwise, the distribution $Q_{XY}^1$ obtained by projecting $Q_{XY}^*$ onto $\mathcal{E}(\theta^\gamma(Q))$ satisfies $Q_{XY}^1 \in C$ and
\begin{equation}
D(Q_{XY}^1 || Q_{XY}) < D(Q_{XY}^* || Q_{XY}),
\end{equation}
(68)which contradict the fact that $Q_{XY}^*$ is the optimizer of (24).

Let $P_{XY}^* \in \mathcal{E}(\theta^\gamma(P))$ be the counterpart of $Q_{XY}^*$ satisfying $Q_X^* = P_X^*$ and $Q_Y^* = P_Y^*$. Then, since
\begin{equation}
D(P_{XY}^* || P_{XY}) = E(Q_{XY}^* || P_{XY}) = r
\end{equation}
(69)and $E(P_{XY}^* || Q_{XY}) = D(Q_{XY}^* || Q_{XY})$, we have $v_1 \geq v_2$.

Let $P_{XY}^*$ be an optimizer of (25), and let $Q_{XY}^*$ be the distribution such that $D(Q_{XY}^* || Q_{XY}) = E(P_{XY}^* || Q_{XY})$ and $Q_X^* = P_X^*$ and $Q_Y^* = P_Y^*$. \footnote{Here, the projection means $Q_{XY}^1$ satisfying $Q_X^1 = Q_X^*$, $Q_Y^1 = Q_Y^*$ and $Q_{XY}^1 \in \mathcal{E}(\theta^\gamma(Q))$.} Since $E(Q_{XY}^* || P_{XY}) \leq D(P_{XY}^* || P_{XY}) \leq r$, we have $v_1 \leq v_2$.

Since $v_1 = v_2$, $Q_{XY}^*$ is also the optimizer of (24). However, since the optimizer of (24) is unique, we have $Q_{XY}^* = Q_{XY}^*$. We can also verify that $P_{XY}^* = P_{XY}^*$ as follows. Note that $F(r)$ is achieved at the boundary, i.e.,
$E(Q_{XY}^*\|P_{XY}) = r$. Since the marginals of $P'_{XY}, Q'_{XY} = Q_{XY}^*$, and $P_{XY}^*$ are the same, and since $P_{XY}^*$ is the unique optimizer of $\min\{D(P_X^*\|P_{XY}): P_X = P_X^*, P_Y = P_Y^*\}$ (see Theorem 1), if $P_{XY}^* \neq P_{XY}^*$, then we have

$$D(P_X^*\|P_{XY}) < D(P_{XY}^*\|P_{XY}),$$

(70)

which contradict the fact that $D(P_{XY}^*\|P_{XY}) = E(Q_{XY}^*\|P_{XY}) = r$ and $D(P_{XY}^*\|P_{XY}) \leq r$. Consequently, the optimizer of (25) is also unique, and it is included in $E(\theta^y(P))$. Finally, the claim that $Q_X^* = P_X^*$ and $Q_Y^* = P_Y^*$ is apparent from the above argument.

By using Lemma 2, we can show the following geometrical properties of the optimizers in (24) and (25), and it plays an important role in the subsequent construction of our new testing scheme.

**Theorem 4** For $0 < r < E(Q_{XY}\|P_{XY})$, the optimization problem $F(r)$ in (24) and (25) is obtained by the unique pair $(Q_X^*, P_{XY}^*) \in E(\theta^y(Q)) \times E(\theta^y(P))$ satisfying the following equations for some $a \in \mathbb{R}\setminus\{0\}, b \in \mathbb{R}$, and and $\lambda = \lambda(r)$:

$$\log \frac{P_{XY}^*(x,y)}{P_{XY}(x,y)} = a \log \frac{Q_{XY}^*(x,y)}{Q_{XY}(x,y)} + b, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y};$$

(71)

$$D(Q_{XY}^*\|Q_{XY}) - D(P_{XY}^*\|P_{XY}) = \lambda,$$

(72)

$$\sum_x P_{XY}^*(x,y) = \sum_x Q_{XY}^*(x,y),$$

(73)

$$\sum_y P_{XY}^*(x,y) = \sum_y Q_{XY}^*(x,y),$$

(74)

and

$$\sum_{x,y} Q_{XY}(x,y) \Lambda_\lambda(x,y) < \lambda < \sum_{x,y} P_{XY}(x,y) \Lambda_\lambda(x,y),$$

(75)

where

$$\Lambda_\lambda(x,y) := \log \frac{Q_{XY}^*(x,y)}{Q_{XY}(x,y)} - \log \frac{P_{XY}^*(x,y)}{P_{XY}(x,y)}. $$

(76)

**Proof**: First, we show that the optimizer pair of (24) and (25) satisfy (71)-(75). From Lemma 2, the optimizer pair of (24) and (25) are given by unique pair $(Q_{XY}^*, P_{XY}^*) \in E(\theta^y(Q)) \times E(\theta^y(P))$ satisfying (73) and (74). Let

$$S_{\geq}(Q_{XY}^*) := \left\{ Q_{XY} : \sum_{x,y} (Q_{XY}(x,y) - Q_{XY}^*(x,y)) \log \frac{Q_{XY}(x,y)}{Q_{XY}^*(x,y)} \geq 0 \right\};$$

(77)

we also define $S_{=} (Q_{XY}^*)$ and $S_{\leq}(Q_{XY}^*)$ by replacing $\geq$ with $=$ and $\geq$ in (77). By Corollary 1, we have $P_{XY}^* \in S_{=} (Q_{XY}^*)$. Since $F(r)$ is achieved at the boundary, i.e., $E(Q_{XY}^*\|P_{XY}) = r$, we have

$$D(P_{XY}^*\|P_{XY}) = E(Q_{XY}^*\|P_{XY}) = r.$$
Furthermore, we also have

$$
\min \{ D(\hat{P}_{XY} \| P_{XY}) : \hat{P}_{XY} \in \mathcal{S}_e(Q_{XY}^\lambda) \} \tag{79}
$$

$$
\geq \inf \{ D(\hat{P}_{XY} \| P_{XY}) : \hat{P}_{XY} \notin \mathcal{S}_e(Q_{XY}^\lambda) \} \tag{80}
$$

$$
\geq \inf \{ D(\hat{P}_{XY} \| P_{XY}) : \hat{P}_{XY} \notin \mathcal{C} \} \tag{81}
$$

$$
\geq \inf \{ E(\hat{P}_{XY} \| P_{XY}) : \hat{P}_{XY} \notin \mathcal{C} \} \tag{82}
$$

where the second inequality follows from $\mathcal{C} \subseteq \mathcal{S}_e(Q_{XY}^\lambda)$ (see (30)). The inequalities (79)-(83) together with (78) imply that $P_{XY}^\lambda$ is the optimizer of (79). Thus, from Theorem 2 $P_{XY}^\lambda$ is contained in the exponential family generated by $P_{XY}$ and $\log Q_{XY}^\lambda(x,y)$, i.e., it can be written as

$$
P_{XY}^\lambda(x,y) = P_{XY}(x,y) \exp \left[ s \log \frac{Q_{XY}^\lambda(x,y)}{Q_{XY}(x,y)} - \psi(s) \right] \tag{84}
$$

for some $s, \psi(s) \in \mathbb{R}$, which implies that the pair $(Q_{XY}^\lambda, P_{XY}^\lambda)$ satisfies (71)\,\,16 The pair also satisfies (72) since $D(Q_{XY}^\lambda \| Q_{XY}) = F(r)$ and (78) (see also (67)). Finally, we can confirm (75) as follows. By noting $P_{XY} \in \mathcal{S}_e(Q_{XY}^\lambda)$ and $D(P_{XY} \| P_{XY}^\lambda) > -D(P_{XY}^\lambda \| P_{XY})$, we have

$$
\sum_{x,y} P_{XY}(x,y) \Lambda(x,y) = \sum_{x,y} P_{XY}(x,y) \left[ \log \frac{Q_{XY}^\lambda(x,y)}{Q_{XY}(x,y)} - \log \frac{P_{XY}^\lambda(x,y)}{P_{XY}(x,y)} \right] \tag{85}
$$

$$
> D(Q_{XY}^\lambda \| Q_{XY}) - D(P_{XY}^\lambda \| P_{XY}) \tag{86}
$$

$$
= \lambda. \tag{87}
$$

Let $\mathcal{S}_e(P_{XY}^\lambda)$ be the set defined by replacing $Q_{XY}$ and $Q_{XY}^\lambda$ with $P_{XY}$ and $P_{XY}^\lambda$ in (77). In fact, by noting (71) and

$$
\sum_{x,y} Q_{XY}^\lambda(x,y) \log \frac{Q_{XY}^\lambda(x,y)}{Q_{XY}(x,y)} = \sum_{x,y} P_{XY}^\lambda(x,y) \log \frac{Q_{XY}^\lambda(x,y)}{Q_{XY}(x,y)}, \tag{88}
$$

which follows from Corollary 1, we can find that either $\mathcal{S}_e(P_{XY}^\lambda) = \mathcal{S}_e(Q_{XY}^\lambda)$ or $\mathcal{S}_e(P_{XY}^\lambda) = \mathcal{S}_e(Q_{XY}^\lambda)$; since $P_{XY} \notin \mathcal{S}_e(P_{XY}^\lambda)$, we have $\mathcal{S}_e(P_{XY}^\lambda) = \mathcal{S}_e(Q_{XY}^\lambda)$. Thus, by noting $Q_{XY} \in \mathcal{S}_e(P_{XY}^\lambda)$ and $-D(Q_{XY} \| Q_{XY}^\lambda) < D(Q_{XY}^\lambda \| Q_{XY})$, we have

$$
\sum_{x,y} Q_{XY}(x,y) \Lambda(x,y) = \sum_{x,y} Q_{XY}(x,y) \left[ \log \frac{Q_{XY}^\lambda(x,y)}{Q_{XY}(x,y)} - \log \frac{P_{XY}^\lambda(x,y)}{P_{XY}(x,y)} \right] \tag{89}
$$

$$
< D(Q_{XY}^\lambda \| Q_{XY}) - D(P_{XY}^\lambda \| P_{XY}) \tag{90}
$$

$$
= \lambda. \tag{91}
$$

Second, for a given pair $(Q_{XY}^\lambda, P_{XY}^\lambda)$ satisfying (71)-(75), we show that the pair is the optimizer pair of (24) and (25); then, since the optimizer pair of (24) and (25) is unique, the solution pair of (71)-(75) is also unique.

16Note that $s \neq 0$ since $P_{XY}^\lambda \neq P_{XY}$. 
We shall show that \( F(D(P_{XY}^\lambda || P_{XY})) = D(Q_{XY}^\lambda || Q_{XY}) \). Then, since \( \lambda(r) \) is one-to-one, \( E(Q_{XY}^\lambda || P_{XY}) = D(P_{XY}^\lambda || P_{XY}) = r \) and \( E(P_{XY}^\lambda || Q_{XY}) = D(Q_{XY}^\lambda || Q_{XY}) = F(r) \), which implies that the pair is the optimizer pair of (24) and (25).

Since \( Q_{XY}^\lambda \) is contained in the set \( C \) of (66) for \( r = D(P_{XY}^\lambda || P_{XY}) \), we have \( D(Q_{XY}^\lambda || Q_{XY}) \geq F(D(P_{XY}^\lambda || P_{XY})) \).

In order to prove the opposite inequality, let \( P_{XY}^* \) be the optimizer of (24) for \( r = D(P_{XY}^\lambda || P_{XY}) \). From Lemma 2 we have \( Q_{XY}^* \in E(\theta_P(Q)) \). Let \( P_{XY}^* \in E(\theta_P(P)) \) be the same marginal counterpart of \( Q_{XY}^* \), i.e., the marginals satisfy \( P_X^* = Q_X^* \) and \( P_Y^* = Q_Y^* \). Note that

\[
D(P_{XY}^* || P_{XY}) = E(Q_{XY}^* || P_{XY}) \leq D(P_{XY}^\lambda || P_{XY})
\]  

(92) since \( Q_{XY}^\lambda \) is contained in the set \( C \) of (66) for \( r = D(P_{XY}^\lambda || P_{XY}) \). Thus, we have

\[
D(P_{XY}^* || P_{XY}) - D(P_{XY}^* || P_{XY}^\lambda) \leq D(P_{XY}^\lambda || P_{XY}).
\]

(93) Furthermore, from Corollary 1 and the fact that \( P_X^* = Q_X^* \) and \( P_Y^* = Q_Y^* \), we have

\[
D(P_{XY}^* || P_{XY}) - D(P_{XY}^* || P_{XY}^\lambda) = \sum_{x,y} Q_{XY}(x,y) \left[ \log \frac{P_{XY}(x,y)}{P_{XY}^*(x,y)} - \log \frac{P_{XY}(x,y)}{P_{XY}^\lambda(x,y)} \right]
\]

(94)

\[
\leq \sum_{x,y} Q_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_{XY}^\lambda(x,y)}.
\]

(95)

(96)

Thus, we have \( Q_{XY}^\lambda \in S_{\geq}(P_{XY}^\lambda) \), where \( S_{\geq}(P_{XY}^\lambda) \) is defined by replacing \( \geq \) with \( \leq \) in the definition of \( S_{\geq}(P_{XY}^\lambda) \).

As we have shown above, (71) and Corollary 1 imply \( S_{\geq}(P_{XY}^\lambda) = S_{\geq}(Q_{XY}^\lambda) \). Thus, we have

\[
F(D(P_{XY}^\lambda || P_{XY})) = D(Q_{XY}^\lambda || Q_{XY})
\]

(97)

\[
\geq \min \{ D(Q_{XY}^\lambda || Q_{XY}) : Q_{XY} \in S_{\geq}(Q_{XY}^\lambda) \}
\]

(98)

\[
= D(Q_{XY}^\lambda || Q_{XY}),
\]

(99)

where the last equality follows since \( Q_{XY}^\lambda \in S_{\geq}(Q_{XY}^\lambda) \) is equivalent to

\[
D(Q_{XY}^\lambda || Q_{XY}) \geq D(Q_{XY}^\lambda || Q_{XY}^\lambda) + D(Q_{XY}^\lambda || Q_{XY}).
\]

(100)

Now, we introduce a proxy of the log-likelihood ratio \( \Lambda(x,y) \) as follows. For \(-E(Q_{XY} || P_{XY}) < \lambda < E(P_{XY} || Q_{XY})\), we define

\[
\Lambda(x,y) := \log \frac{Q_{XY}(x,y)}{Q_{XY}(x,y)} - \log \frac{P_{XY}(x,y)}{P_{XY}(x,y)}
\]

(101)

for the unique solution pair \( (Q_{XY}^\lambda, P_{XY}^\lambda) \) satisfying (71)-(75); for \( \lambda = E(P_{XY} || Q_{XY}) \), we define

\[
\Lambda(x,y) := \log \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}
\]

(102)
where $P_{XY}^*$ is the optimizer of $E(P_{XY}||Q_{XY})$; for $\lambda = -E(Q_{XY}||P_{XY})$, we define
\[
\Lambda_\lambda(x, y) := -\log \frac{Q_{XY}^*(x, y)}{P_{XY}(x, y)},
\]
(103)
where $Q_{XY}^*$ is the optimizer of $E(Q_{XY}||P_{XY})$.

Since $(Q_{XY}^\lambda, P_{XY}^\lambda) \in \mathcal{E}(\theta^\eta(Q)) \times \mathcal{E}(\theta^\eta(P))$\footnote{Note also that $P_{XY}^\lambda \in \mathcal{E}(\theta^\eta(Q))$ and $Q_{XY}^\lambda \in \mathcal{E}(\theta^\eta(P))$.}, i.e., the pairs $Q_{XY}^\lambda$ and $P_{XY}^\lambda$ have the same values as $Q_{XY}$ and $P_{XY}$ at the $\theta^\eta$-coordinate, respectively, we can write
\[
\Lambda_\lambda(x, y) = a_1(x) + a_2(y)
\]
(104)
for some functions $a_1$ on $\mathcal{X}$ and $a_2$ on $\mathcal{Y}$. Thus, for any joint distribution $P_{\tilde{X}\tilde{Y}} \in \mathcal{F}(\mathcal{X} \times \mathcal{Y})$ with marginals $P_{\tilde{X}}$ and $P_{\tilde{Y}}$, it holds that
\[
\sum_{x, y} P_{\tilde{X}}(x)P_{\tilde{Y}}(y)\Lambda_\lambda(x, y) = \sum_x P_{\tilde{X}}(x)a_1(x) + \sum_y P_{\tilde{Y}}(y)a_2(y)
\]
(105)
\[
= \sum_{x, y} P_{\tilde{X}\tilde{Y}}(x, y)\Lambda_\lambda(x, y).
\]
(106)

More generally, $\sum_{x, y} \tilde{P}_{\tilde{X}\tilde{Y}}(x, y)\Lambda_\lambda(x, y)$ and $\sum_{z, y} P_{\tilde{X}\tilde{Y}}(x, y)\Lambda_\lambda(x, y)$ take the same value as long as the marginals of $\tilde{P}_{\tilde{X}\tilde{Y}}$ and $P_{\tilde{X}\tilde{Y}}$ coincide. Geometrically, this is because $\mathcal{M}(\eta^x, \eta^y)$ for a given $(\eta^x, \eta^y)$ is orthogonal to $\mathcal{E}(\theta^\eta(P))$ and $\mathcal{E}(\theta^\eta(Q))$.

Remark 1 Functions $a_1, a_2$ in decomposition \cite{104} are not unique; for instance, by some calculation, we can verify (see Appendix A) that $\Lambda_\lambda(x, y)$ can be decomposed as\footnote{Note that $\mathcal{X} = \{0, 1, \ldots, |\mathcal{X}| - 1\}$ and $\mathcal{Y} = \{0, 1, \ldots, |\mathcal{Y}| - 1\}$. In fact, the choice $(0, 0)$ is not crucial, and the same statement holds even if we replace $\Lambda_\lambda(x, 0), \Lambda_\lambda(0, y)$, and $\Lambda_\lambda(0, 0)$ with $\Lambda_\lambda(x, j), \Lambda_\lambda(i, y)$, and $\Lambda_\lambda(i, j)$ for arbitrarily fixed $(i, j) \in \mathcal{X} \times \mathcal{Y}$, respectively.}
\[
\Lambda_\lambda(x, y) = \Lambda_\lambda(x, 0) + \Lambda_\lambda(0, y) - \Lambda_\lambda(0, 0)
\]
(107)

Now, we are ready to propose our testing scheme. Fix arbitrary $-E(Q_{XY}||P_{XY}) \leq \lambda \leq E(P_{XY}||Q_{XY})$. Upon observing $x$ and $y$, the encoders send their types, i.e., $f_1(x) = t_x$ and $f_2(y) = t_y$, respectively. Then, upon receiving a pair of marginal types $(t_x, t_y)$, the decoder $g$ outputs $H_0$ if
\[
\sum_{x, y} t_x(x)t_y(y)\Lambda_\lambda(x, y) > \tau
\]
(108)
holds for a prescribed threshold $\tau$; otherwise it outputs $H_1$. By \cite{106}, the decoder outputs $g(f_1(x), f_2(y)) = H_0$ if and only if
\[
\sum_{x, y} t_{xy}(x, y)\Lambda_\lambda(x, y) = \frac{1}{n} \sum_{i=1}^n \Lambda_\lambda(x_i, y_i)
\]
(109)
\[
> \tau.
\]
(110)

Apparently, this scheme is a symmetric scheme. The performance of this scheme is summarized in the following theorem.
Fig. 2: (a) A description of Han-Kobayashi’s testing scheme; and descriptions of our proposed scheme with (b) \( \lambda = E(P_{XY} \| Q_{XY}) \), (c) \( \lambda = -E(Q_{XY} \| P_{XY}) \), and (d) \(-E(Q_{XY} \| P_{XY}) < \lambda < E(P_{XY} \| Q_{XY}). \)

**Theorem 5** For a given \(-E(Q_{XY} \| P_{XY}) \leq \lambda \leq E(P_{XY} \| Q_{XY})\) and \( \tau \in \mathbb{R} \), the above mentioned Neyman-Pearson-like scheme \( T_n^{\text{NPl}} \) has the following error trade-off:

\[
\alpha[T_n^{\text{NPl}}] = P\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_{\lambda}(X_i, Y_i) \leq \tau\right),
\]

\[
\beta[T_n^{\text{NPl}}] = Q\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_{\lambda}(X_i, Y_i) > \tau\right).
\]

Although it is not clear whether the Neyman-Pearson-like test is optimal or not for a given blocklength, we will numerically examine that the Neyman-Pearson-like test has better error trade-off than the Hoeffding-like test in Section \[V\] Furthermore, for the large-deviation regime and the second-order regime, we will show that the Neyman-Pearson-like test is asymptotically optimal in Section \[VI\] and Section \[VII\] respectively.
We close this section by comparing Han-Kobayashi’s scheme (see Section II-C) and the Neyman-Pearson-like scheme proposed above. In Han-Kobayashi’s scheme, a type \( P_{XY} \) is accepted if and only if there exists some \( \bar{P}_{XY} \) satisfying \( P_X = \bar{P}_X, P_Y = \bar{P}_Y, \) and \( D(\bar{P}_{XY} \| P_{XY}) < r. \) In other words, a type is accepted if and only if it is included in the cylinder of radius \( r \) given by

\[
\bigcup_{\eta: D(P_{XY,\alpha} \| P_{XY}) < r} \mathcal{M}(\eta^*, \eta^r)
\]

(see [17]); see Fig. 2a.

For a given \( -E(Q_{XY} \| P_{XY}) \leq \lambda \leq E(Q_{XY} \| P_{XY}) \), let

\[
Q^\lambda_{XY,s}(x, y) := Q_{XY}(x, y) \exp\{s \Lambda_\lambda(x, y) - \psi_{Q,\lambda}(s)\},
\]

\[
\psi_{Q,\lambda}(s) := \log \sum_{x,y} Q_{XY}(x, y) \exp\{s \Lambda_\lambda(x, y)\},
\]

and let

\[
\mathcal{E}_\lambda(Q) := \{Q^\lambda_{XY,s} : s \in \mathbb{R}\}
\]

be the exponential family containing \( Q_{XY} \). Let

\[
P^\lambda_{XY,t}(x, y) := P_{XY}(x, y) \exp\{t \Lambda_\lambda(x, y) - \psi_{P,\lambda}(t)\},
\]

\[
\psi_{P,\lambda}(t) := \log \sum_{x,y} P_{XY}(x, y) \exp\{t \Lambda_\lambda(x, y)\},
\]

and let

\[
\mathcal{E}_\lambda(P) := \{P^\lambda_{XY,t} : t \in \mathbb{R}\}
\]

be the exponential family containing \( P_{XY} \). For \( \lambda = \lambda := E(P_{XY} \| Q_{XY}) \), note that \( \mathcal{E}_\lambda(Q) \) is the e-geodesic connecting \( Q_{XY} \) and \( P^\lambda_{XY} \); for \( \lambda = \lambda := -E(Q_{XY} \| P_{XY}) \), note that \( \mathcal{E}_\lambda(P) \) is the e-geodesic connecting \( P_{XY} \) and \( Q^\lambda_{XY} \).

For \( \tau \in \mathbb{R} \), let

\[
\mathcal{M}_\lambda(\tau) := \left\{P_{XY} : \sum_{x,y} P_{XY}(x, y) \Lambda_\lambda(x, y) = \tau\right\}
\]

be the mixture family generated by \( \Lambda_\lambda(x, y) \).

In our Neyman-Pearson-like testing scheme with \( \lambda = \lambda = E(P_{XY} \| Q_{XY}) \), we bisect the entire space with \( \mathcal{M}_\lambda(\tau) \) that is orthogonal to \( \mathcal{E}_\lambda(Q) \), the e-geodesic connecting \( Q_{XY} \) and \( P^\lambda_{XY} \); see Fig. 2b. On the other hand, in our Neyman-Pearson testing scheme with \( \lambda = -E(Q_{XY} \| P_{XY}) \), we bisect the entire space with \( \mathcal{M}_\lambda(\tau) \) that is orthogonal to \( \mathcal{E}_\lambda(P) \), the e-geodesic connecting \( P_{XY} \) and \( Q^\lambda_{XY} \); see Fig. 2c. Furthermore, for \( -E(Q_{XY} \| P_{XY}) < \lambda < E(P_{XY} \| Q_{XY}) \), we bisect the entire space with \( \mathcal{M}_\lambda(\tau) \) that is orthogonal to neither \( \mathcal{E}_\lambda(Q) \) nor \( \mathcal{E}_\lambda(P) \); see Fig. 2d. In fact, the mixture family \( \mathcal{M}_\lambda(\tau) \) is orthogonal to \( \mathcal{E}_\lambda(Q) \) and \( \mathcal{E}_\lambda(P) \). In particular, when \( \tau = \lambda \), the intersections are given by \( Q^\lambda_{XY} \) and \( P^\lambda_{XY} \), respectively.

---

\(^{19}\)Here, “orthogonal” means that the Pythagorean theorem with respect to the relative entropy holds at the intersection.
In contrast to the standard hypothesis testing problem, our Neyman-Pearson-like testing scheme has the freedom to choose the direction of bisection with parameter $\lambda$. In fact, as we determine in later sections, an appropriate choice of $\lambda$ depending on a target threshold $\tau$ is very important.

**Remark 2** By using the coordinate notation, we can rewrite (see Appendix B) the condition in (71) as
\[
\theta_i^x(P_{XY}^\lambda) - \theta_i^x(P_{XY}) = a[\theta_i^x(Q_{XY}^\lambda) - \theta_i^x(Q_{XY})], \quad \forall i = 1, \ldots, d_x,
\]
(121)
\[
\theta_j^y(P_{XY}^\lambda) - \theta_j^y(P_{XY}) = a[\theta_j^y(Q_{XY}^\lambda) - \theta_j^y(Q_{XY})], \quad \forall j = 1, \ldots, d_y.
\]
(122)
Thus, we can find that the e-geodesic connecting $P_{XY}$ and $P_{XY}^\lambda$ and that connecting $Q_{XY}$ and $Q_{XY}^\lambda$ are parallel each other (see also Fig. 4 in Section V).

**Remark 3** When $\theta^x_y(P) = \theta^x_y(Q)$, we have $P_{XY}^\lambda = Q_{XY}^\lambda$, which implies
\[
\Lambda_\lambda(x, y) = \log \frac{P_{XY}(x, y)}{Q_{XY}(x, y)}.
\]
(123)
Thus, our Neyman-Pearson-like testing scheme reduces to the Neyman-Pearson testing scheme of the standard hypothesis testing between $P_{XY}$ and $Q_{XY}$. In other words, when the correlation components of $P_{XY}$ and $Q_{XY}$ are the same and only the marginals are different, then a test based on zero-rate encoding is as good as a test which is based on full-rate encoding.

**Remark 4** By noting
\[
\sum_{x,y} P_{XY}^\lambda(x, y) \Lambda_\lambda(x, y) = \sum_{x,y} Q_{XY}^\lambda(x, y) \Lambda_\lambda(x, y) = \lambda
\]
(124)
and (71), we can rewrite the condition
\[
\sum_{x,y} P_{XY}(x, y) \Lambda_\lambda(x, y) = \lambda
\]
(125)
as
\[
\sum_{x,y} P_{XY}(x, y) \log \frac{Q_{XY}^\lambda(x, y)}{Q_{XY}(x, y)} = D(Q_{XY}^\lambda \parallel Q_{XY})
\]
(126)
or
\[
\sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}^\lambda(x, y)}{P_{XY}(x, y)} = D(P_{XY}^\lambda \parallel P_{XY}).
\]
(127)
Thus, the mixture family plane $M_\lambda(\lambda)$ is the tangent plane of the cylinder given by (113) (see also (77) and (29)).

**Remark 5** There is a trivial testing scheme that is most powerful among the class of symmetric schemes. Upon observing marginal types $(P_X, P_Y)$, the decoder accept the null hypothesis if
\[
\log \frac{P(t_{X^n} = P_X, t_{Y^n} = P_Y)}{Q(t_{X^n} = P_X, t_{Y^n} = P_Y)} > \tau
\]
(128)
for a prescribed threshold $\tau$. In fact, this is the Neyman-Pearson test such that the pair of marginal types $(t_{X^n}, t_{Y^n})$ is regarded as the observation. However, since computing the log-likelihood ratio in (128) is intractable as the
blocklength becomes larger, it is difficult to implement this scheme. On the other hand, in the Neyman-Pearson-like scheme introduced above, we only need to compute the empirical average of $\Lambda_X(x, y)$ with respect to the product of marginal types, and it is easier to implement.

V. Example

In this section, we consider the binary example, i.e., $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and compare the error trade-off of our proposed testing scheme and Han-Kobayashi’s scheme. In the binary case, $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is parametrized by three parameters $(\theta^x, \theta^y, \theta^{xy})$ or $(\eta^x, \eta^y, \eta^{xy})$. For simplicity of notation, we denote $\tilde{\theta} = \theta(P)$ and $\tilde{\theta} = \theta(Q)$; similar notations are used for the expectation parameters. For given $(\tilde{\eta}^x, \tilde{\eta}^y)$ and $\tilde{\theta}^{xy}$, by noting (58), the intersection of $\mathcal{E}(\tilde{\theta}^{xy})$ and $\mathcal{M}(\tilde{\eta}^x, \tilde{\eta}^y)$ can be derived by solving the following equation with respect to $\eta^{xy}$:

$$\log \frac{\eta^{xy} (1 - \tilde{\eta}^x - \tilde{\eta}^y + \eta^{xy})}{(\tilde{\eta}^x - \eta^{xy})(\tilde{\eta}^y - \eta^{xy})} = \tilde{\theta}^{xy}. \quad (129)$$

The above equation is equivalent to

$$(e^{\tilde{\theta}^{xy}} - 1)(\eta^{xy})^2 - [(\tilde{\eta}^x + \tilde{\eta}^y)(e^{\tilde{\theta}^{xy}} - 1) + 1]\eta^{xy} + e^{\tilde{\theta}^{xy}} \tilde{\eta}^x \tilde{\eta}^y. \quad (130)$$

When $\tilde{\theta}^{xy} = 0$, i.e., there is no correlation, then $\eta^{xy} = \tilde{\eta}^x \tilde{\eta}^y$ is the only solution. When $\tilde{\theta}^{xy} \neq 0$, we have two solutions:

$$\eta^{xy} = \frac{[(\tilde{\eta}^x + \tilde{\eta}^y)(e^{\tilde{\theta}^{xy}} - 1) + 1] \pm \sqrt{[(\tilde{\eta}^x + \tilde{\eta}^y)(e^{\tilde{\theta}^{xy}} - 1) + 1]^2 - 4(e^{\tilde{\theta}^{xy}} - 1)e^{\tilde{\theta}^{xy}} \tilde{\eta}^x \tilde{\eta}^y}}{2(e^{\tilde{\theta}^{xy}} - 1)}. \quad (131)$$

Here, note that $\eta^{xy}$ must satisfy

$$\eta^{xy} \leq \min[\tilde{\eta}^x, \tilde{\eta}^y] \quad (132)$$

$$\leq \frac{\tilde{\eta}^x + \tilde{\eta}^y}{2}. \quad (133)$$

Thus, when $\tilde{\theta}^{xy} > 0$, then

$$\eta^{xy} = \frac{[(\tilde{\eta}^x + \tilde{\eta}^y)(e^{\tilde{\theta}^{xy}} - 1) + 1] - \sqrt{[(\tilde{\eta}^x + \tilde{\eta}^y)(e^{\tilde{\theta}^{xy}} - 1) + 1]^2 - 4(e^{\tilde{\theta}^{xy}} - 1)e^{\tilde{\theta}^{xy}} \tilde{\eta}^x \tilde{\eta}^y}}{2(e^{\tilde{\theta}^{xy}} - 1)} \quad (134)$$

is the only valid solution since the other solution violates (133). On the other hand, note that $\eta^{xy}$ is required to be nonnegative. Thus, when $\tilde{\theta}^{xy} < 0$, then (134) is the only valid solution in this case too since the other solution is negative.

The use of (134) enables us to numerically solve (71)-(75) to find $(Q_{\mathcal{X}Y}^\lambda, P_{\mathcal{X}Y}^\lambda)$ for each $\lambda$. More specifically, for given parameters $(\eta^x, \eta^y)$, by using (134), we can compute $(\hat{Q}_{\mathcal{X}Y}, \hat{P}_{\mathcal{X}Y}) \in \mathcal{E}(\tilde{\theta}^{xy}(Q)) \times \mathcal{E}(\tilde{\theta}^{xy}(P))$ such that the marginals satisfy $\hat{Q}_X(1) = \hat{P}_X(1) = \eta^x$ and $\hat{Q}_Y(1) = \hat{P}_Y(1) = \eta^y$. Then, we can numerically solve (72), (121), and (122) with respect to $(\eta^x, \eta^y)$ to find $(Q_{\mathcal{X}Y}^\lambda, P_{\mathcal{X}Y}^\lambda)$. \footnote{For this binary example, we were able to find the solution by using builtin functions of Mathematica.} Fig. 3 compares the trade-off between the type I error rate and the type II error rate for each $\lambda$. More specifically, for given parameters $(\eta^x, \eta^y)$, by using (134), we can compute $(\hat{Q}_{\mathcal{X}Y}, \hat{P}_{\mathcal{X}Y}) \in \mathcal{E}(\tilde{\theta}^{xy}(Q)) \times \mathcal{E}(\tilde{\theta}^{xy}(P))$ such that the marginals satisfy $\hat{Q}_X(1) = \hat{P}_X(1) = \eta^x$ and $\hat{Q}_Y(1) = \hat{P}_Y(1) = \eta^y$. Then, we can numerically solve (72), (121), and (122) with respect to $(\eta^x, \eta^y)$ to find $(Q_{\mathcal{X}Y}^\lambda, P_{\mathcal{X}Y}^\lambda)$. \footnote{For this binary example, we were able to find the solution by using builtin functions of Mathematica.} Fig. 3 compares the trade-off between the type I error rate and the type II error rate for each $\lambda$. More specifically, for given parameters $(\eta^x, \eta^y)$, by using (134), we can compute $(\hat{Q}_{\mathcal{X}Y}, \hat{P}_{\mathcal{X}Y}) \in \mathcal{E}(\tilde{\theta}^{xy}(Q)) \times \mathcal{E}(\tilde{\theta}^{xy}(P))$ such that the marginals satisfy $\hat{Q}_X(1) = \hat{P}_X(1) = \eta^x$ and $\hat{Q}_Y(1) = \hat{P}_Y(1) = \eta^y$. Then, we can numerically solve (72), (121), and (122) with respect to $(\eta^x, \eta^y)$ to find $(Q_{\mathcal{X}Y}^\lambda, P_{\mathcal{X}Y}^\lambda)$. \footnote{For this binary example, we were able to find the solution by using builtin functions of Mathematica.}
Fig. 3: A comparison of the trade-off between the type I error probability (horizontal axis) and the type II error probability (vertical axis) for our proposed testing scheme (red dots) and Han-Kobayashi’s scheme (blue dots). The distributions $P_{XY}$ and $Q_{XY}$ are chosen to be \((135)\), and the block length is $n = 100$.  

error probability and the type II error probability for our proposed testing scheme with $\tau = \lambda$ and Han-Kobayashi’s scheme (see Theorem 5 and Proposition 2). The distributions are chosen to be  

$$
P_{XY} = \begin{bmatrix} 1/2 & 1/8 \\
1/8 & 1/4 \end{bmatrix}, \quad Q_{XY} = \begin{bmatrix} 1/8 & 1/4 \\
1/2 & 1/8 \end{bmatrix}, \tag{135}
$$

and the block length is $n = 100$. The plots in the figure indicate that, for a short block length such as $n = 100$, our proposed testing scheme outperforms the previously known scheme of Han-Kobayashi.  

In Fig. 4, we plotted the trajectories of \((\theta(x(P_{XY}^\lambda)), \theta(y(P_{XY}^\lambda)))\) and \((\theta(x(Q_{XY}^\lambda)), \theta(y(Q_{XY}^\lambda)))\) by varying $\lambda$, where the distributions are the same as \((135)\). For visual convenience, we inserted vectors \((\theta(x(P_{XY}^\lambda) - \theta(x(P_{XY})), \theta(y(P_{XY}^\lambda) - \theta(y(P_{XY})))\) and \((\theta(x(Q_{XY}^\lambda) - \theta(x(Q_{XY})), \theta(y(Q_{XY}^\lambda) - \theta(y(Q_{XY})))\) for two values of $\lambda$ (purple vectors and green vectors, respectively). As is predicted from \((71)\) (see also Remark 2), we can verify that the vectors of the same color are parallel to each other.  

VI. LARGE DEVIATION REGIME  

In this section, we discuss the large deviation performance of our proposed testing scheme. For this purpose, we need some preparations. Recall the notations introduced at the end of Section IV. Since the potential functions $\psi_{Q,\lambda}(s)$ and $\psi_{P,\lambda}(t)$ of the exponential families are strict convex, $\psi'_{Q,\lambda}(s)$ and $\psi'_{P,\lambda}(t)$ are monotonically increasing  

$^{21}$Numerically computing the exact trade-off for multinomial distribution becomes computationally intractable as the block length becomes larger.
functions. Thus, we can define the inverse functions $s_{Q,\lambda}(\tau)$ and $t_{P,\lambda}(\tau)$ of $\psi'_{Q,\lambda}(s)$ and $\psi'_{P,\lambda}(t)$ by

$$\psi'_{Q,\lambda}(s_{Q,\lambda}(\tau)) = \sum_{x,y} Q_{XY,s_{Q,\lambda}(\tau)}(x,y)\Lambda_{\lambda}(x,y)$$

$$= \tau,$$  \hspace{1cm} (136)

$$\psi'_{P,\lambda}(t_{P,\lambda}(\tau)) = \sum_{x,y} P_{XY,t_{P,\lambda}(\tau)}(x,y)\Lambda_{\lambda}(x,y)$$

$$= \tau,$$  \hspace{1cm} (138)

respectively. Note that $s_{Q,\lambda}(\tau)$ and $t_{P,\lambda}(\tau)$ are the expectation parameters of the exponential families $\mathcal{E}_{\lambda}(Q)$ and $\mathcal{E}_{\lambda}(P)$. We also use the following expressions of the relative entropies:

$$D(Q_{XY,s_{Q,\lambda}(\tau)}||Q_{XY}) = s_{Q,\lambda}(\tau)\tau - \psi_{Q,\lambda}(s_{Q,\lambda}(\tau)),$$  \hspace{1cm} (140)

$$D(P_{XY,t_{P,\lambda}(\tau)}||P_{XY}) = t_{P,\lambda}(\tau)\tau - \psi_{P,\lambda}(t_{P,\lambda}(\tau)).$$  \hspace{1cm} (141)

**Theorem 6** For $-E(Q_{XY}||P_{XY}) < \lambda < E(P_{XY}||Q_{XY})$, the Neyman-Pearson-like testing scheme $T_{n}^{\text{NPI}}$ of Section [IV] with threshold $\tau = \lambda$ satisfies

$$\alpha[T_{n}^{\text{NPI}}] \leq \exp\{-nD(P_{XY}^{\lambda}||P_{XY})\},$$  \hspace{1cm} (142)

$$\beta[T_{n}^{\text{NPI}}] \leq \exp\{-nD(Q_{XY}^{\lambda}||Q_{XY})\}$$

$$= \exp\{-nF(D(P_{XY}^{\lambda}||P_{XY}))\}$$  \hspace{1cm} (144)

**Proof:** By applying the Markov inequality to Theorem 5 for any $t \leq 0$ and $s \geq 0$, we have (see [143])

$$P\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_{\lambda}(X_{i},Y_{i}) < \tau\right) \leq \exp\{-n(t\tau - \psi_{P,\lambda}(t))\}$$  \hspace{1cm} (145)
and
\[ Q \left( \frac{1}{n} \sum_{i=1}^{n} \Lambda_{\lambda}(X_i, Y_i) \geq \tau \right) \leq \exp \{-n(s\tau - \psi_{Q,\lambda}(s))\}. \] (146)

From (137) and (75), we have
\[ \psi'_{Q,\lambda}(s_{Q,\lambda}(\lambda)) = \sum_{x,y} Q_{XY,s_{Q,\lambda}(\lambda)}(x,y) \Lambda_{\lambda}(x,y) \]
\[ = \lambda \] (147)
\[ > \sum_{x,y} Q_{XY}(x,y) \Lambda_{\lambda}(x,y) \] (148)
\[ = \sum_{x,y} Q_{XY,0}(x,y) \Lambda_{\lambda}(x,y) \] (149)
\[ = \psi'_{Q,\lambda}(0), \] (150)

which, together with the fact that \( \psi'_{Q,\lambda}(s) \) is an increasing function, imply \( s_{Q,\lambda}(\lambda) > 0 \). Similarly, from (139) and (75), we have \( t_{P,\lambda}(\lambda) < 0 \).

From (72) and Corollary 1 combined with the same marginal conditions (see (73) and (74)), we have
\[ \lambda = D(Q_{XY}^\lambda \| Q_{XY}) - D(P_{XY}^\lambda \| P_{XY}) \]
\[ = \sum_{x,y} Q_{XY}^\lambda(x,y) \left[ \log \frac{Q_{XY}^\lambda(x,y)}{Q_{XY}(x,y)} - \log \frac{P_{XY}^\lambda(x,y)}{P_{XY}(x,y)} \right] \]
\[ = \sum_{x,y} Q_{XY}^\lambda(x,y) \Lambda_{\lambda}(x,y). \] (154)

Furthermore, (71) together with the definition of \( Q_{XY,s}^\lambda \) (see (114)) implies \( Q_{XY}^\lambda \in \mathcal{E}_\lambda(Q) \). This together with (154) means that \( Q_{XY}^\lambda = Q_{XY,s_{Q,\lambda}(\lambda)}^\lambda \). Similarly, we have \( P_{XY}^\lambda = P_{XY,t_{P,\lambda}(\lambda)}^\lambda \).

Consequently, by substituting \( \tau = \lambda \), \( t = t_{P,\lambda}(\lambda) \), and \( s = s_{Q,\lambda}(\lambda) \) into (145) and (146), we have (142) and (143) from (140) and (141). Furthermore, by (67) and (72) for \( r = D(P_{XY}^\lambda \| P_{XY}) \), we have (144).

Note that Theorem 6 means that our Neyman-Pearson-like testing scheme with threshold \( \tau = \lambda \) is optimal in the large deviation regime. Compared to the derivation of the same exponents based on the method of type, Theorem 6 has the advantage in that there is no polynomial factor of \( n \) that stems from the number of types.

In Fig. 5, we plotted the trade-off of the two exponents
\[ (D(P_{XY}^\lambda \| P_{XY}), F(D(P_{XY}^\lambda \| P_{XY}))) = (D(P_{XY}^\lambda \| P_{XY}), D(Q_{XY}^\lambda \| Q_{XY})) \] (155)
by varying \( \lambda \in [-E(Q_{XY} \| P_{XY}), E(P_{XY} \| Q_{XY})] \). For fixed values of \( \lambda \), say \( \lambda = \bar{\lambda} = E(P_{XY} \| Q_{XY}) \) or \( \lambda = \underline{\lambda} = -E(Q_{XY} \| P_{XY}) \), we can also achieve the following trade-offs by using our testing scheme \(^{22}\)
\[ (D(P_{XY,t_{P,\lambda}(\tau)}^\lambda \| P_{XY}), D(Q_{XY,s_{Q,\lambda}(\tau)}^\lambda \| Q_{XY})) \] (156)

\(^{22}\)To prove the achievability of (155) (or (157)), we take \( \lambda = \bar{\lambda} \) (or \( \lambda = \underline{\lambda} \)), \( t = t_{P,\lambda}(\tau) \), and \( s = s_{Q,\lambda}(\tau) \) in (145) and (146).
Fig. 5: A comparison of the trade-offs between the type I and type II exponent, where the horizontal axis is type I exponent and the vertical axis is type II exponent. The red solid curve is the optimal trade-off between the type I exponent and the type II exponent, i.e., (155); the blue dashed curve is the trade-off between the type I exponent and type II exponent for $\lambda = \lambda$, i.e., (156); the green dotted curve is the trade-off between the type I exponent and the type II exponent for $\lambda = \lambda$, i.e., (157). The distributions $P_{XY}$ and $Q_{XY}$ are chosen to be (135).

for $\sum_{x,y} Q_{XY}(x,y) \Lambda_\tau(x,y) \leq \tau \leq E(P_{XY} \| Q_{XY})$ and

$$D(P_{XY,\tau} \| P_{XY}), D(Q_{XY,\tau} \| Q_{XY})$$

(157)

for $-E(Q_{XY} \| P_{XY}) \leq \tau \leq \sum_{x,y} P_{XY}(x,y) \Lambda_\tau(x,y)$. For comparison, we also plotted these trade-offs in Fig. 5.

The adjustment of $\lambda$ is crucial to achieve the optimal trade-off; when either the type I exponent or the type II exponent is very small, then $\lambda = \lambda$ or $\lambda = \lambda$ is quite effective.

VII. Second-Order Analysis

Let us start by introducing the projected relative entropy density:

$$j_{P \| Q}(x,y) := \log \frac{P_{XY}(x,y)}{Q_{XY}(x,y)},$$

(158)

where $P_{XY}$ is the optimizer of $E(P_{XY} \| Q_{XY})$. From Corollary 1 we have the following proposition, which justifies referring to $j_{P \| Q}$ as the projected relative entropy density.

Proposition 4 It holds that

$$E[j_{P \| Q}(X,Y)] = E(P_{XY} \| Q_{XY}),$$

(159)

where the expectation is taken over $(X,Y) \sim P_{XY}$.

Proposition 1 shows that the optimal exponent of the type II error probability is given by $E(P_{XY} \| Q_{XY})$, and it can be achieved by a symmetric scheme. In this section, for a given constraint $0 < \epsilon < 1$ on the type I error...
probability, we consider the optimal second-order exponent among the class of symmetric schemes:

\[ G_s(\varepsilon) := \sup \left\{ \liminf_{n \to \infty} \frac{-\log \beta[T_n] - nE(P_{XY} \| Q_{XY})}{\sqrt{n}} : \{T_n\}_{n=1}^\infty \text{ is symmetric, } \limsup_{n \to \infty} \alpha[T_n] \leq \varepsilon \right\}. \] (160)

In approximate terms, this means that the type I error probability and type II error probability behave as follows:

\[ \alpha[T_n] \simeq \varepsilon, \quad \beta[T_n] \simeq \exp\{-nE(P_{XY} \| Q_{XY}) - \sqrt{n}G_s(\varepsilon)\}. \] (161)

The following theorem characterizes the second-order coefficient \( G_s(\varepsilon) \).

**Theorem 7** For a given \( 0 < \varepsilon < 1 \), it holds that

\[ G_s(\varepsilon) = \sqrt{V(P \| Q)}\Phi^{-1}(\varepsilon), \] (162)

where

\[ V(P \| Q) := \text{Var}[J_{P \| Q}(X,Y)] \] (163)

for \((X,Y) \sim P_{XY}\).

**Remark 6** In fact, we can present a slightly stronger statement than Theorem 7 for the achievability, i.e., the Neyman-Pearson-like testing scheme \( T_{n^{\text{NP1}}} \) of Section IV performs as follows for sufficiently large \( n \):

\[ \alpha[T_{n^{\text{NP1}}}] \leq \varepsilon \] (164)

and

\[ -\log \beta[T_{n^{\text{NP1}}}] \geq nE(P_{XY} \| Q_{XY}) + \sqrt{nV(P \| Q)}\Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + O(1). \] (165)

**A. Proof of Achievability of Theorem 7**

We prove the stronger statement, i.e., Remark 6. We use the following technical lemma shown in [28, Lemma 47].

**Lemma 3** Let \( Z_1, \ldots, Z_n \) be i.i.d. random variables with \( \sigma^2 = \text{Var}[Z_i] \), \( T = E[|Z_i - E[Z_i]|^3] \). Then, for any \( \gamma \),

\[ \mathbb{E} \left[ \exp \left( -\sum_{i=1}^{n} Z_i \right) \mathbb{1}\left\{ \sum_{i=1}^{n} Z_i > \gamma \right\} \right] \leq 2 \left( \frac{\log 2}{\sqrt{2\pi}} + \frac{12T}{\sigma^2} \right) \frac{\exp(-\gamma)}{\sigma \sqrt{n}}, \] (166)

where \( \mathbb{1}\{\cdot\} \) is the indicator function.
From Theorem 5 with $\lambda = E(P_{XY}||Q_{XY})$ the testing scheme of Section 4 satisfies

$$\beta[T_n^{SP}] = Q \left( \frac{1}{n} \sum_{i=1}^{n} J_{P||Q}(X_i, Y_i) > \tau \right)$$

(167)

$$= \sum_{x,y} Q_{X,Y}^n(x, y) 1 \left\{ \log \frac{P_{X,Y}^{n}(x, y)}{Q_{X,Y}^{n}(x, y)} > \tau n \right\}$$

(168)

$$= \sum_{x,y} P_{X,Y}^{n}(x, y) \exp \left( - \log \frac{P_{X,Y}^{n}(x, y)}{Q_{X,Y}^{n}(x, y)} \right) 1 \left\{ \log \frac{P_{X,Y}^{n}(x, y)}{Q_{X,Y}^{n}(x, y)} > \tau n \right\}$$

(169)

$$\leq 2 \left( \frac{\log 2 + 12T}{\sigma^2} \right) \exp \left( - \frac{\tau n}{\sigma \sqrt{n}} \right),$$

(170)

where we used Lemma 3 by setting $Z_i = J_{P||Q}(X_i, Y_i)$ for $(X_i, Y_i) \sim P_{XY}$ in the last inequality. Now, we set

$$\tau = E(P_{XY}||Q_{XY}) + \sqrt{\frac{V(P||Q)}{n}} \Phi^{-1} \left( \varepsilon - \frac{6T(P||Q)}{\sqrt{nV(P||Q)^{3/2}}} \right),$$

(171)

where $T(P||Q)$ is the absolute third moment of $J_{P||Q}(X, Y)$ for $(X, Y) \sim P_{XY}$. Then, application of the Berry-Esseen theorem (see [15]) yields

$$\alpha[T_n^{SP}] = P \left( \frac{1}{n} \sum_{i=1}^{n} J_{P||Q}(X_i, Y_i) \leq \tau \right)$$

(172)

$$\leq \varepsilon,$$

(173)

On the other hand, (170), (171), and the Taylor approximation of $\Phi^{-1}(\cdot)$ around $\varepsilon$ yields

$$- \log \beta[T_n^{SP}] \geq \tau n + \log(\sigma \sqrt{n}) + O(1)$$

(174)

$$= nE(P_{XY}||Q_{XY}) + \sqrt{nV(P||Q)} \Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + O(1),$$

(175)

which completes the achievability proof.

B. Proof of Converse of Theorem 7

First, we provide a simple converse bound for the class of symmetric schemes. For a given joint type $P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$, let

$$E_n(P_{XY}||Q_{XY}) := \min \left\{ D(\tilde{P}_{XY}||Q_{XY}) : \tilde{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}), \tilde{P}_X = P_X, \tilde{P}_Y = P_Y \right\}.$$

(176)

By definition, $E_n(P_{XY}||Q_{XY}) \geq E(P_{XY}||Q_{XY})$ for any $P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$.

**Proposition 5** For any $r > 0$ and for any symmetric scheme $T_n$ such that

$$\beta[T_n] \leq \exp \{ -rn \},$$

(177)

it holds that

$$\alpha[T_n] \geq P \left( E_n(t_{X^nY^n}||Q_{XY}) < r - (|\mathcal{X}||\mathcal{Y}| - 1) \frac{\log(n+1)}{n} \right).$$

(178)

23Note that $\Lambda(x, y) = J_{P||Q}(x, y)$ for $\lambda = E(P_{XY}||Q_{XY})$. August 17, 2017 DRAFT
Proof: Since $T_n$ is a symmetric scheme, without loss of generality, we can assume that the decoder $g$ only depends on marginal types $(P_{\bar{X}}, \bar{P}_Y) \in \mathcal{P}_n(X) \times \mathcal{P}_n(Y)$. We claim that, if
\[
E_n(P_{\bar{X}} \times P_{\bar{Y}} \parallel Q_{XY}) < r - (|X||Y| - 1) \frac{\log(n+1)}{n},
\] (179)
then $g(P_{\bar{X}}, \bar{P}_Y) = H_1$. Otherwise, by letting $P_{\bar{X}} \in \mathcal{P}_n(X \times Y)$ be such that $P_{\bar{X}} = \bar{P}_X, \bar{P}_Y = P_{\bar{Y}},$ and $E_n(P_{\bar{X}} \times P_{\bar{Y}} \parallel Q_{XY}) = D(P_{\bar{X}} \parallel Q_{XY})$, we have
\[
\beta[T_n] \geq Q_{XY}(T_{\bar{X}\bar{Y}})
\] (180)
\[
\geq \frac{1}{(n+1)(|X||Y|-1)} \exp\{-nD(P_{\bar{X}} \parallel Q_{XY})\}
\] (181)
\[
= \frac{1}{(n+1)(|X||Y|-1)} \exp\{-nE_n(P_{\bar{X}} \times P_{\bar{Y}} \parallel Q_{XY})\}
\] (182)
\[
> \exp\{-rn\},
\] (183)
which contradict (177), where $T_{\bar{X}\bar{Y}}$ is the set of all pairs $(x, y)$ such that $t_{xy} = P_{\bar{X}}$. Thus, we have (178). □

Second, we approximate $E_n(P_{\bar{X}} \parallel Q_{XY})$ by $E(P_{\bar{X}} \parallel Q_{XY})$.

**Lemma 4** For any $P_{\bar{X}} \in \mathcal{P}_n(X \times Y)$, it holds that
\[
E_n(P_{\bar{X}} \parallel Q_{XY}) \leq E(P_{\bar{X}} \parallel Q_{XY}) + \Delta_n,
\] (184)
where
\[
\Delta_n := \nu_n \log \frac{|X||Y|}{\nu_n} + \nu_n \max_{x,y} \log \frac{1}{Q_{XY}(x,y)}.
\] (185)

and
\[

\nu_n := 4(|X| - 1)(|Y| - 1)
\] (186)

Proof: Let $P_{\bar{X}}^* \in \mathcal{P}_n(X \times Y)$ be such that $P_{\bar{X}}^* = P_{\bar{X}}, P_{\bar{Y}}^* = P_{\bar{Y}}$, and
\[
D(P_{\bar{X}}^* \parallel Q_{XY}) = E(P_{\bar{X}}^* \parallel Q_{XY}).
\] (187)

We set $\tilde{P}_{XY} \in \mathcal{P}_n(X \times Y)$ as follows: $\tilde{P}_X = P_{\bar{X}}^*, \tilde{P}_Y = P_{\bar{Y}}^*$, and
\[
\tilde{P}_{XY}(x, y) = \left[ P_{\bar{X}}^*(x, y) \right]_{1/n}
\] (188)
for $x = 1, \ldots, |X| - 1$ and $y = 1, \ldots, |Y| - 1$, where
\[
|t|_{1/n} := \max \left\{ \frac{k}{n} : k \in \mathbb{N}, \frac{k}{n} \leq t \right\}.
\] (189)
Then, we have
\[ \| \hat{P}_{XY} - P_{XY}^* \| \]
\[ = \left[ \hat{P}_X(0) + \hat{P}_Y(0) - 1 + \sum_{x=1}^{\mid X \mid - 1} \sum_{y=1}^{\mid Y \mid - 1} \hat{P}_{XY}(x, y) \right] - \left[ P_{XY}^*(0) + P_{XY}^*(0) - 1 + \sum_{x=1}^{\mid X \mid - 1} \sum_{y=1}^{\mid Y \mid - 1} P_{XY}^*(x, y) \right] \]
\[ + \sum_{x=1}^{\mid X \mid - 1} \left[ \hat{P}_X(x) - \sum_{y=1}^{\mid Y \mid - 1} \hat{P}_{XY}(x, y) \right] - \left[ P_{XY}^*(x) - \sum_{y=1}^{\mid Y \mid - 1} P_{XY}^*(x, y) \right] \]
\[ + \sum_{y=1}^{\mid Y \mid - 1} \left[ \hat{P}_Y(y) - \sum_{x=1}^{\mid X \mid - 1} \hat{P}_{XY}(x, y) \right] - \left[ P_{XY}^*(y) - \sum_{x=1}^{\mid X \mid - 1} P_{XY}^*(x, y) \right] \]
\[ + \sum_{x=1}^{\mid X \mid - 1} \sum_{y=1}^{\mid Y \mid - 1} \left| \hat{P}_{XY}(x, y) - P_{XY}^*(x, y) \right| \]
\[ \leq \nu_n. \]

Thus, by the continuity of the entropy [12 Lemma 2.7], we have
\[ E_n(P_{XY} \| Q_{XY}) \leq D(\hat{P}_{XY} \| Q_{XY}) \]
\[ = -H(\hat{P}_{XY}) + \sum_{x,y} \hat{P}_{XY}(x,y) \log \frac{1}{Q_{XY}(x,y)} \]
\[ \leq -H(P_{XY}^*) + \nu_n \log \frac{|X||Y|}{\nu_n} + \sum_{x,y} P_{XY}^*(x,y) \log \frac{1}{Q_{XY}(x,y)} + \nu_n \max_{x,y} \log \frac{1}{Q_{XY}(x,y)} \]
\[ = D(P_{XY}^* \| Q_{XY}) + \Delta_n. \]

Next, we show a useful relationship between the projected relative entropy \( E(P_{XY} \| Q_{XY}) \) and its density \( J_P \| Q \).

**Lemma 5** Let \( \tilde{\eta} \) be the expectation parameter of \( P_{XY} \) (see Section III). Then, it holds that
\[ \frac{\partial E(P_{XY,\eta} \| Q_{XY})}{\partial \eta_i} \bigg|_{\eta=\tilde{\eta}} = J_P \| Q(\tilde{i},0) - J_P \| Q(0,0), \]
\[ \frac{\partial E(P_{XY,\eta} \| Q_{XY})}{\partial \eta_j} \bigg|_{\eta=\tilde{\eta}} = J_P \| Q(0,\tilde{j}) - J_P \| Q(0,0), \]
\[ \frac{\partial E(P_{XY,\eta} \| Q_{XY})}{\partial \eta_{ij}} \bigg|_{\eta=\tilde{\eta}} = 0 \]
for \( 1 \leq i \leq d_x \) and \( 1 \leq j \leq d_y \), where \( \tilde{\eta} = \eta(P) \) is the expectation parameter of \( P_{XY} \).

**Proof:** We first show (202). In fact, this follows from the fact that
\[ E(P_{XY,\tilde{\eta}+\Delta} \| Q_{XY}) = E(P_{XY,\tilde{\eta}} \| Q_{XY}) \]
for every vector \( \Delta \) such that the values are 0 except for the \( \eta_{ij} \)-coordinate.

We prove (200) by first showing that
\[ \frac{\partial E(P_{XY,\tilde{\eta}} \| Q_{XY})}{\partial \eta_i} \bigg|_{\eta=\tilde{\eta}} = \frac{\partial E(P_{XY,\tilde{\eta}} \| Q_{XY})}{\partial \eta_i} \bigg|_{\eta=\tilde{\eta}}, \]
where \( \hat{\eta} \) is the expectation parameter of \( P_{XY}^* \). Since the \((\eta^x, \eta^y)\)-coordinate of \( \hat{\eta} \) and \( \hat{\eta} \) are the same, note that

\[
E(P_{XY,\hat{\eta}+\Delta}\|Q_{XY}) = E(P_{XY,\hat{\eta}}\|Q_{XY})
\] (205)

holds for every vector \( \Delta \) such that the values are 0 with the exception of the \( \eta^* \)-coordinate. Let \( \Delta \eta^* \) be the \( \eta^* \)-coordinate of \( \Delta \). Then, we have

\[
\frac{\partial E(P_{XY,\eta}\|Q_{XY})}{\partial \eta^*} \bigg|_{\eta=\hat{\eta}} = \frac{\lim_{\Delta \eta^* \to 0} E(P_{XY,\eta+\Delta}\|Q_{XY}) - E(P_{XY,\hat{\eta}}\|Q_{XY})}{\Delta \eta^*}
\] (206)

Thus, we have

\[
\frac{\partial E(P_{XY,\eta}\|Q_{XY})}{\partial \eta^*} \bigg|_{\eta=\hat{\eta}} = \frac{\lim_{\Delta \eta^* \to 0} E(P_{XY,\eta+\Delta}\|Q_{XY}) - E(P_{XY,\hat{\eta}}\|Q_{XY})}{\Delta \eta^*} = \frac{\partial E(P_{XY,\eta}\|Q_{XY})}{\partial \eta^*} \bigg|_{\eta=\hat{\eta}}.
\] (207)

Now, note that (see Proposition 4)

\[
E(P_{XY,\eta}\|Q_{XY}) = \sum_{x,y} P_{XY,\eta}(x, y) \log \frac{P_{XY,\eta}(x, y)}{Q_{XY}(x, y)},
\] (209)

where \( P_{XY,\eta}^* \) is the optimizer of \( E(P_{XY,\eta}\|Q_{XY}) \). Furthermore, from (210), we have

\[
\frac{\partial P_{XY,\eta}(x, y)}{\partial \eta^*} = \left( \delta_i(x) - \delta_0(x) \right) \delta_0(y).
\] (210)

Thus, we have

\[
\frac{\partial E(P_{XY,\eta}\|Q_{XY})}{\partial \eta^*} \bigg|_{\eta=\hat{\eta}} = \sum_{x,y} \frac{\partial P_{XY,\eta}(x, y)}{\partial \eta^*} \bigg|_{\eta=\hat{\eta}} \log \frac{P_{XY,\eta}(x, y)}{Q_{XY}(x, y)}
\] (211)

\[
+ \sum_{x,y} P_{XY,\eta}(x, y) \frac{\partial}{\partial \eta^*} \left[ \log \frac{P_{XY,\eta}(x, y)}{Q_{XY}(x, y)} \right] \bigg|_{\eta=\hat{\eta}}
\] (212)

\[
= \log \frac{P_{XY,\eta}^*(i, 0)}{Q_{XY}(i, 0)} - \log \frac{P_{XY,\eta}^*(0, 0)}{Q_{XY}(0, 0)}
\] (213)

\[
+ \sum_{x,y} \frac{\partial}{\partial \eta^*} P_{XY,\eta}(x, y) \bigg|_{\eta=\hat{\eta}}
\] (214)

\[
= \log \frac{P_{XY,\eta}^*(i, 0)}{Q_{XY}(i, 0)} - \log \frac{P_{XY,\eta}^*(0, 0)}{Q_{XY}(0, 0)}
\] (215)

\[
= J_{P\|Q}(i, 0) - J_{P\|Q}(0, 0),
\] (216)

where the second equality follows from (210) and \( P_{XY,\eta}^* = P_{XY,\hat{\eta}} \), the third equality follows from \( \frac{\partial}{\partial \eta^*} 1 = 0 \), and the last equality follows from the fact that \( P_{XY,\hat{\eta}}^* = P_{XY}^* \). Finally, (201) is proved in a similar manner. ■

From Lemma 5 we have the following approximation of \( E(P_{XY}\|Q_{XY}) \) around \( P_{XY} \).

**Theorem 8** It holds that

\[
E(P_{XY}\|Q_{XY}) = \sum_{x,y} P_{XY}(x, y) J_{P\|Q}(x, y) + o(P_{XY} - P_{XY}).
\] (217)

August 17, 2017 DRAFT
Proof: From Lemma [5] by noting that \((\eta^x, \eta^y)\)-coordinates correspond to the marginal distributions, the first-order Taylor approximation of \(E(P_{XY}\|Q_{XY})\) around \(P_{XY}\) is given by

\[
E(P_{XY}\|Q_{XY}) = E(P_{XY}\|Q_{XY}) + \sum_{i=1}^{d_x} (P_X(i) - P_X(i)) (j_{P\|Q}(i,0) - j_{P\|Q}(0,0)) + \sum_{j=1}^{d_y} (P_Y(j) - P_Y(j)) (j_{P\|Q}(0,j) - j_{P\|Q}(0,0)) + o(||P_{XY} - P_{XY}||) \tag{218}
\]

Thus, by using Theorem 8 for \(E(P_{XY}\|Q_{XY})\) around \(P_{XY}\) is given by

\[
E(P_{XY}\|Q_{XY}) = E(P_{XY}\|Q_{XY}) + \sum_{x} (P_X(x) - P_X(x)) j_{P\|Q}(x,0) + \sum_{y} (P_Y(y) - P_Y(y)) j_{P\|Q}(0,y) + o(||P_{XY} - P_{XY}||) \tag{219}
\]

\[
\alpha[T_n] \geq P\left(E(t_{X^Y^n}\|Q_{XY}) < E(P_{XY}\|Q_{XY}) + \frac{G}{\sqrt{n}}\right). \tag{225}
\]

Let

\[
K_n := \left\{ P_{XY} \in \mathcal{P}_{n}(\mathcal{X} \times \mathcal{Y}) : |P_{XY}(x,y) - P_{XY}(x,y)| \leq \sqrt{\frac{\log n}{n}} \forall (x,y) \right\}, \tag{226}
\]

Then, by the Hoeffding inequality, we have

\[
P\left(t_{X^Y^n} \notin K_n\right) \leq \frac{2|\mathcal{X}||\mathcal{Y}|}{n^2}. \tag{227}
\]

Thus, by using Theorem [8] for \(P_{XY} \in K_n\), we have

\[
\alpha[T_n] \geq P\left(t_{X^Y^n} \in K_n, E(t_{X^Y^n}\|Q_{XY}) < E(P_{XY}\|Q_{XY}) + \frac{G}{\sqrt{n}}\right) \tag{228}
\]

\[
\geq P\left(t_{X^Y^n} \in K_n, \sum_{x,y} t_{X^Y^n}(x,y) j_{P\|Q}(x,y) < E(P_{XY}\|Q_{XY}) + \frac{G}{\sqrt{n}} - c\frac{\log n}{n}\right) \tag{229}
\]

\[
\geq P\left(\frac{1}{n} \sum_{i=1}^{n} j_{P\|Q}(X_i, Y_i) < E(P_{XY}\|Q_{XY}) + \frac{G}{\sqrt{n}} - c\frac{\log n}{n}\right) - P\left(t_{X^Y^n} \notin K_n\right) \tag{230}
\]

\[
\geq P\left(\frac{1}{n} \sum_{i=1}^{n} j_{P\|Q}(X_i, Y_i) < E(P_{XY}\|Q_{XY}) + \frac{G}{\sqrt{n}} - c\frac{\log n}{n}\right) - \frac{2|\mathcal{X}||\mathcal{Y}|}{n^2}. \tag{231}
\]
for some constant $c > 0$. Thus, by the central limit theorem, we have

$$\liminf_{n \to \infty} \alpha [T_n] > \varepsilon,$$

which together with (223) implies

$$G_s(\varepsilon) \leq \sqrt{V(P||Q)} \Phi^{-1}(\varepsilon).$$

(VIII. CONCLUSION AND DISCUSSION)

In this paper, we proposed a novel testing scheme for the zero-rate multiterminal hypothesis testing problem. The previously known scheme of Han-Kobayashi is based on a cylinder with respect to the relative entropy, and thus their scheme can be regarded as a multiterminal analogue of the Hoeffding test. In contrast, our proposed scheme bisects the joint probability simplex by an appropriate mixture family, an approach reminiscent of the Neyman-Pearson test. For a short block length, we numerically determined that the performance of our proposed scheme is superior to that of the previously reported scheme. We also showed that, in a large deviation regime, our proposed scheme optimizes the trade-off of exponents that was shown by Han-Kobayashi. Furthermore, we derived the optimal second-order exponent among the class of symmetric schemes, which can be achieved by our proposed scheme.

More ambitious goals would be to identify the optimal second-order exponent or to derive non-asymptotic bounds for the class of general zero-rate testing schemes. However, such analyses are not only technically difficult, but they also have subtlety in the problem formulations. For instance, a straightforward definition of the optimal second-order exponent by replacing “symmetric” with “zero-rate” in (160) would make no sense. In fact, suppose that the marginals of $P_{XY}$ and $Q_{XY}$ are the same as an extreme case. In that case, it is known that the optimal first-order exponent achievable by zero-rate schemes is 0. However, a trivial zero-rate scheme allows the encoders to send the first $\lceil n^\gamma \rceil$ symbols of their observations for a given $\frac{1}{2} < \gamma < 1$; then the type II error probability behaves as $\exp\{-cn^\gamma\}$ for some constant $c > 0$. In other words, the order of the second-order rate may depend on the growth rate of the message sizes even if they are zero-rate. Identifying appropriate formulations and studying these problems would be important in future.

APPENDIX

A. Proof of Remark [7]

We only prove the statement for $-E(Q_{XY}||P_{XY}) < \lambda < E(P_{XY}||Q_{XY})$: the two extreme cases follow by a similar argument. Since $Q^\lambda_{XY} \in \mathcal{E}(\theta_{\lambda}(Q))$, it has the same values as $Q_{XY}$ at the $\theta_{\lambda}$-coordinate. Thus, it holds that (see (38)), for $x \neq 0$ and $y \neq 0$,

$$\log \frac{Q^\lambda_{XY}(x,0)Q^\lambda_{XY}(0,y)Q_{XY}(x,y)Q_{XY}(0,0)}{Q^\lambda_{XY}(x,y)Q^\lambda_{XY}(0,0)Q_{XY}(x,0)Q_{XY}(0,y)} = 0.$$  (234)
By noting this fact, we have
\[
\log \frac{Q_X^\lambda(x,0)}{Q_X(x,0)} + \log \frac{Q_Y^\lambda(0,y)}{Q_Y(0,y)}
\]
\[= \log \frac{Q_X^\lambda(x,0)Q_Y^\lambda(y,0)}{Q_X(x,0)Q_Y(0,y)}
\]
\[= \log \frac{Q_X^\lambda(x,y)Q_Y^\lambda(0,0)}{Q_X(x,y)Q_Y(0,0)Q_X(x,0)Q_Y(0,y)}
\]
\[= \log \frac{Q_X^\lambda(x,y)}{Q_X(x,y)} + \log \frac{Q_Y^\lambda(0,0)}{Q_Y(0,0)}
\]
(235)
(236)
(237)
(238)
(239)
for \(x \neq 0\) and \(y \neq 0\); also, the identity (235)-(239) trivially holds when either \(x = 0\) or \(y = 0\). Similarly, since \(P_X^\lambda \in \mathcal{E}(\theta^\psi(P))\), we have
\[
\log \frac{P_X^\lambda(x,0)}{P_X(x,0)} + \log \frac{P_Y^\lambda(0,y)}{P_Y(0,y)} = \log \frac{P_X^\lambda(x,y)}{P_X(x,y)} + \log \frac{P_Y^\lambda(0,0)}{P_Y(0,0)},
\]
(240)
which together with the identity (235)-(239) imply
\[
A_\lambda(x,0) + A_\lambda(0,y) = A_\lambda(x,y) + A_\lambda(0,0).
\]
(241)
B. Proof of Remark 2

By noting that \(\theta^\psi\) components of the pair \(P_X^\lambda\) and \(P_Y\) and the pair \(Q_X^\lambda\) and \(Q_Y\) are the same, respectively, and by noting (35), we can rewrite (71) as
\[
\sum_{i=1}^{d_x} [\theta_i^x(P_X^\lambda) - \theta_i^x(P_X)] \delta_i(x) + \sum_{j=1}^{d_y} [\theta_j^y(P_Y^\lambda) - \theta_j^y(P_Y)] \delta_j(y) - \psi(\theta(P_X^\lambda)) + \psi(\theta(P_X))
\]
\[= a \sum_{i=1}^{d_x} [\theta_i^x(Q_X^\lambda) - \theta_i^x(Q_X)] \delta_i(x) + a \sum_{j=1}^{d_y} [\theta_j^y(Q_Y^\lambda) - \theta_j^y(Q_Y)] \delta_j(y) - a\psi(\theta(Q_X^\lambda)) + a\psi(\theta(Q_X)) + b
\]
(242)
(243)
for every \((x, y) \in \mathcal{X} \times \mathcal{Y}\). By substituting \((x, y) = (0, 0)\), we have
\[
-\psi(\theta(P_X^\lambda)) + \psi(\theta(P_X)) = -a\psi(\theta(Q_X^\lambda)) + a\psi(\theta(Q_X)) + b.
\]
(244)
Thus, we have
\[
\sum_{i=1}^{d_x} [\theta_i^x(P_X^\lambda) - \theta_i^x(P_X)] \delta_i(x) + \sum_{j=1}^{d_y} [\theta_j^y(P_Y^\lambda) - \theta_j^y(P_Y)] \delta_j(y)
\]
\[= a \sum_{i=1}^{d_x} [\theta_i^x(Q_X^\lambda) - \theta_i^x(Q_X)] \delta_i(x) + a \sum_{j=1}^{d_y} [\theta_j^y(P_Y^\lambda) - \theta_j^y(P_Y)] \delta_j(y)
\]
(245)
(246)
which implies (121) and (122) since \(\delta_i(x)\) and \(\delta_j(y)\) for \(i = 1, \ldots, d_x\) and \(j = 1, \ldots, d_y\) are linearly independent as functions on \(\mathcal{X} \times \mathcal{Y}\).
ACKNOWLEDGEMENT

The author would like to thank Hiroshi Nagaoka for helpful suggestions on information geometry and encouragement, Shigeaki Kuzuoka for fruitful discussion, and Po-Ning Chen for comments on an earlier version of the manuscript. The author is supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 16H06091.

REFERENCES

[1] R. Ahlswede and I. Csiszár, “Hypothesis testing with communication constraints,” IEEE Trans. Inform. Theory, vol. 32, no. 4, pp. 533–542, July 1986.
[2] S. Amari, “Fisher information under restriction of Shannon information in multi-terminal situations,” Annals of the Institute of Statistical Mathematics, vol. 41, no. 4, pp. 623–648, 1989.
[3] ——, “Information geometry on hierarchy of probability distributions,” IEEE Trans. Inform. Theory, vol. 47, no. 5, pp. 1701–1711, July 2001.
[4] ——, “On optimal data compression in multiterminal statistical inference,” IEEE Trans. Inform. Theory, vol. 57, no. 9, pp. 5577–5587, September 2011.
[5] ——, Information Geometry and Its Applications. Springer, 2016.
[6] S. Amari and T. S. Han, “Statistical inference under multiterminal rate restrictions: A differential geometric approach,” IEEE Trans. Inform. Theory, vol. 35, no. 2, pp. 217–227, March 1989.
[7] S. Amari and H. Nagaoka, Methods of Information Geometry. Oxford University Press, 2000.
[8] T. Berger, “Decentralized estimation and decision theory,” in Presented at the IEEE 7th Spring Workshop on Information Theory, Mt. Kisco, NY, September 1979.
[9] R. E. Blahut, “Hypothesis testing and information theory,” IEEE Trans. Inform. Theory, vol. 20, no. 4, pp. 405–417, July 1974.
[10] I. Csiszár, “I-Divergence geometry of probability distributions and minimization problems,” The Annals of Probability, vol. 3, no. 1, pp. 146–158, 1975.
[11] ——, “The method of types,” IEEE Trans. Inform. Theory, vol. 44, no. 6, pp. 2505–2523, October 1998.
[12] I. Csiszár and J. Körner, Information Theory, Coding Theorems for Discrete Memoryless Systems, 2nd ed. Cambridge University Press, 2011.
[13] I. Csiszár and P. C. Shields, Information Theory and Statistics: A Tutorial. Now Publishers Inc., 2004.
[14] A. Dembo and O. Zeitouni, Large Deviations Techniques and Applications, 2nd ed. Springer, 1998.
[15] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 2. Wiley, 1971.
[16] T. S. Han, “Hypothesis testing with multiterminal data compression,” IEEE Trans. Inform. Theory, vol. 33, no. 6, pp. 759–772, November 1987.
[17] T. S. Han and S. Amari, “Statistical inference under multiterminal data compression,” IEEE Trans. Inform. Theory, vol. 44, no. 6, pp. 2300–2324, October 1998.
[18] T. S. Han and K. Kobayashi, “Exponential-type error probabilities for multiterminal hypothesis testing,” IEEE Trans. Inform. Theory, vol. 35, no. 1, pp. 2–14, January 1989.
[19] T. S. Han and S. Verdú, “Approximation theory of output statistics,” IEEE Trans. Inform. Theory, vol. 39, no. 3, pp. 752–772, May 1993.
[20] M. Hayashi, “Information spectrum approach to second-order coding rate in channel coding,” IEEE Trans. Inform. Theory, vol. 55, no. 11, pp. 4947–4966, November 2009.
[21] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” The Annals of Mathematical Statistics, vol. 36, no. 2, pp. 369–401, April 1965.
[22] A. Ingber and Y. Kochman, “The dispersion of lossy source coding,” in Proc. Data Compression Conference, 2011, pp. 53–62.
[23] V. Kostina and S. Verdú, “Fixed-length lossy compression in the finite blocklength regime,” IEEE Trans. Inform. Theory, vol. 58, no. 6, pp. 3309–3338, June 2012.
[24] E. L. Lehmann and J. P. Romano, Testing Statistical Hypotheses, 3rd ed. Springer, 2005.
[25] K. Nakagawa and F. Kanaya, “On the converse theorem in statistical hypothesis testing,” *IEEE Trans. Inform. Theory*, vol. 39, no. 2, pp. 623–628, March 1993.

[26] J. Neyman and E. S. Pearson, “On the use and interpretation of certain test criterion for purposes of statistical inference,” *Biometrika*, vol. 20A, pp. 175–240, 264–299, 1928.

[27] Y. Polyanskiy, “Hypothesis testing via a comparator,” in *Proc. IEEE Int. Symp. Inf. Theory 2012*, Cambridge, MA, 2012, pp. 2206–2210.

[28] Y. Polyanskiy, H. V. Poor, and S. Verdu, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[29] H. M. H. Shalaby and A. Papamarcou, “Multiterminal detection with zero-rate data compression,” *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 254–267, March 1992.

[30] V. Strassen, “Asymptotische Abschätzungen in Shannons Informationstheorie,” in *Trans. Third. Prague Conf. Inf. Th.*, 1962, pp. 689–723.

[31] M. Ueta and S. Kuzuoka, “The error exponent of zero-rate multiterminal hypothesis testing for sources with common information,” in *Proc. ISITA 2014*, Melbourne, Australia, 2014, pp. 559–563.

[32] S. Watanabe, “Second-order region for Gray-Wyner network,” *IEEE Trans. Inform. Theory*, vol. 63, no. 2, pp. 1006–1018, February 2017.

[33] Z. Zhang and T. Berger, “Estimation via compressed information,” *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 198–211, March 1988.

[34] L. Zhou, V. Y. F. Tan, and M. Motani, “Discrete lossy Gray-Wyner revisited: Second-order asymptotics, large and moderate deviations,” *IEEE Trans. Inform. Theory*, vol. 63, no. 3, pp. 1766–1791, March 2017.