Distributed Testing With Cascaded Encoders

Wenwen Zhao, *Student Member, IEEE,* and Lifeng Lai, *Member, IEEE*

***Abstract*—In this paper, we consider distributed testing problems with cascaded encoders, which allow cascaded communications among encoders so that each encoder can utilize messages from other encoders for encoding. We first focus on a special case of testing against independence and design a scheme that enables each encoder to take advantage of extra information from other encoders. We also derive a matching upper bound and prove that the designed scheme is optimal. We then investigate the case with general hypotheses and obtain a lower bound on the type 2 error exponent. We further compare the performances that can be achieved by schemes with and without cascaded communications. We show that cascaded communication improves the performance in terms of the type 2 error exponent under positive rate communication constraints. On the other hand, we prove that cascaded communication does not provide performance gain under zero-rate communication constraints.**

***Index Terms*—Distributed learning, error exponent, hypothesis testing, cascaded communications.**

I. INTRODUCTION

Recently, there have been growing interests in distributed inference and machine learning problems [2]–[8], in which available data is distributed over multiple terminals connected by links with limited communication capacity. These problems are mainly motivated by the explosive growth in the size and scale of modern datasets such that data is either naturally collected by multiple terminals or is too large to be fitted into one terminal. The limited communication budget between the terminals makes the inference problems more challenging than those in the centralized scenario.

Moreover, due to the importance of communication in the distributed scenario, methods under two different communication styles: namely non-interactive communication and interactive communication, are under active investigation. In the non-interactive communication case, each terminal uses only its own data to determine the communication message to be sent. In this case, each terminal compresses its own observations and sends compressed data to the decision maker, who will then perform inference and make a decision based on its own data along with compressed data received from terminals [5], [6], [9]–[17]. In the interactive communication cases, each terminal can utilize messages received from other terminals along with its own data to determine the message to be sent. The scenario with two fully interactive encoders is also under active research [7], [18], [19]. However, the scenarios where multiple encoders interact with each other are still open.

As the scenario with general interactive communication among multiple encoders is very complex and difficult to allow meaningful progress, in this paper, we study a special form of encoder interaction from information theoretic perspective. In particular, we consider statistical inference with multiple cascaded encoders. In the considered setup, there are L terminals (encoders) $\mathcal{X}_l$, $l = 1, \cdots, L$ and a decision terminal $\mathcal{Y}$, in which terminal $\mathcal{X}_l$ has data related to random variable X_l only. All random variables $(X_1, \cdots X_L, Y)$ take values in a finite set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L \times \mathcal{Y}$. We consider a special form of interaction among terminals, where we assume that terminals broadcast their messages in a sequential order from terminal 1 until terminal L, and each terminal uses all messages received so far along with its own observations for encoding. More specifically, terminal $\mathcal{X}_1$ first broadcasts its encoded message based on its own observations X_1^n, and then terminal $\mathcal{X}_2$ broadcasts its encoded message based on both its own observations X_2^n and the message received from terminal $\mathcal{X}_1$. This process continues until terminal $\mathcal{X}_L$ broadcasts its message based on its own observations X_L^n and all messages received before. Finally, terminal $\mathcal{Y}$ performs statistical inference based on messages received from terminals $\mathcal{X}_l$, $l = 1, \cdots, L$ and its own data related to Y. This model is motivated by scenarios where the encoders may not send their messages to the decision maker at the same time, and hence the encoder who sends message at later time can take advantage of the messages overheard from other encoder's transmission (e.g., when these messages are transmitted over wireless channels). In this paper, we focus on a basic inference problem in which terminal $\mathcal{Y}$ tries to decide the joint probability mass function (PMF) of the data from the following two hypotheses:

$$H_0 : P_{X_1 \cdots X_L Y} \ \text{ vs } \ H_1 : Q_{X_1 \cdots X_L Y}.$$

Our goal is to maximize the type 2 error exponent under constraints on the type 1 error probability and communication rates. We note that one can consider other forms of interactions among encoders. However, the problem becomes very complex if an arbitrary form of interaction among encoders are allowed. These cases are left for future study.

We first focus on the problem of testing against independence, in which $Q_{X_1 \cdots X_L Y} = P_{X_1 \cdots X_L} P_Y$ and hence we are interested in determining whether $(X_1, \cdots, X_L)$ and Y are independent or not. This work builds upon our recent work [14], in which we studied the non-interactive communication case under the same hypotheses. Compared with [14], this paper allows cascaded communication for terminals $\mathcal{X}_l, l = 1, \cdots, L$,

The authors are with Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616. Email:{wwzhao, lflai}@ucdavis.edu. This paper was presented in part in IEEE Information Theory Workshop, 2015 [1]. The work was supported by National Science Foundation under Grants CNS-1660128 and CCF-1717943.

so that terminal $\mathcal{X}_l$ can utilize the information from terminals $\mathcal{X}_{l'}, l' = 1, \cdots, l-1$, when it performs encoding. The cascaded communication results in two major differences with the cases using non-interactive communication [9], [14]. First, in the non-interactive communication case, one typically converts the testing against independence problem to the problem of source coding with a helper [9], then uses the corresponding results in the source coding with a helper problem to characterize the type 2 error exponent. However, if we follow a similar strategy, then the problem will be related to a source coding with multiple helpers problem, which is still an open problem in network information theory. Second, in the existing work with multiple terminals under non-interactive communication as studied in [14], the type 2 error exponent is not fully characterized. However, in our cascaded communication case, as terminals are allowed to use the received massages to perform encoding, we are able to fully characterize the type 2 error exponent for certain scenarios.

We then extend the study to the case with general hypotheses. The problem with non-interactive communication under the same hypotheses was first proposed and studied in [9] and a tighter lower bound was derived in [11]. Different from these works, in which it is assumed that data related to all $X_l, l = 1, \cdots, L$ is stored in one terminal $\mathcal{X}$ (and hence there are two terminals $\mathcal{X}$ and $\mathcal{Y}$ in the model studied in [11] and [9]), we allow data related to $X_l, l = 1, \cdots, L$ to be stored in multiple terminals, and we allow cascaded communications among encoders for encoding. As these two extensions make this problem more complex and no upper bound is derived even for the case with non-interactive communications, in this paper, we only give a lower bound on the type 2 error exponent given the constraints on the type 1 error probability and communication rates.

Finally, we compare performances of schemes with cascaded and non-interactive communications. Intuitively, compared with the scheme with non-interactive communication in [14], the decision maker can potentially obtain more information in the cascaded communication case and hence is expected to make a better decision. We show that this is indeed the case by giving an explicit example in which our scheme with cascaded communication achieves a larger type 2 error exponent under different communication rate constraints. On the other hand, we prove that, compared with non-interactive communication, cascaded communication does not offer any improvement in the type 2 error exponent for the zero-rate data compression case.

The problem studied in this paper is related to but different from several existing interesting works on distributed hypothesis testing problems with interactive communications [7], [9], [19], [20]. In particular, in [7], [19], the authors discussed the case in which $X_l, l = 1, \cdots, L$ are all at terminal $\mathcal{X}$ (and hence $(X_1, \cdots, X_L)$ can be denoted as one random variable X) and terminal $\mathcal{X}$ and terminal $\mathcal{Y}$ can communicate with each other in multiple rounds. [20] considered the same setup with [19] but used sample-by-sample processing, i.e. scaler quantization at each stage. Different from these interesting studies, in our problem, we consider a case in which $X_l, l = 1, \cdots, L$ are at different terminals. Furthermore, cascaded communication among the encoders $\mathcal{X}_l$s is allowed and vector quantization is applied at each stage.

The remainder of the paper is organized as follows. In Section II, we introduce the model studied in this paper. In Section III, we study the problem of testing against independence. In Section III-C, we give a lower bound on the type 2 error exponent for the general case. In Section IV, we compare the performances of schemes with cascaded communication and schemes with non-interactive communication. Finally, we offer some concluding remarks in Section V.

II. Model

In this section, we present our model and summarize the difference between our model and the existing work with non-interactive communication schemes.

A. Model

Consider a set of random variables $(X_1, \cdots X_L, Y)$ taking values in a finite set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L \times \mathcal{Y}$ and admitting a joint PMF that has two possible forms:

$$\begin{aligned} H_0 : &\quad P_{X_1 \cdots X_L Y}, \\ H_1 : &\quad Q_{X_1 \cdots X_L Y}. \end{aligned} \tag{1}$$

$(X_1^n, \cdots X_L^n, Y^n)$ are independently and identically generated according to one of the above joint PMFs, and are observed at different terminals. In particular, terminal $\mathcal{X}_i$ observes only X_i^n and terminal $\mathcal{Y}$ observes only Y^n. We consider a model in which these terminals broadcast messages in a sequential order from terminal 1 until terminal L, and each terminal will use all messages received so far along with its own observations for encoding. More specifically, terminal $\mathcal{X}_1$ will first broadcast its encoded message, which depends only on X_1^n, and then terminal $\mathcal{X}_2$ will broadcast its encoded message, which now depends on not only its own observations X_2^n but also the message received from terminal $\mathcal{X}_1$. The process continues until terminal $\mathcal{X}_L$, who will use messages received from $\mathcal{X}_1$ until $\mathcal{X}_{L-1}$ and its own observations X_L^n for encoding. Finally, terminal $\mathcal{Y}$ decides which hypothesis is true based on its own information and the received messages from terminal $\mathcal{X}_1, \cdots, \mathcal{X}_L$. The system model is illustrated in Fig. 1. In the following, we will use the term "decision maker" and terminal $\mathcal{Y}$ interchangeably.

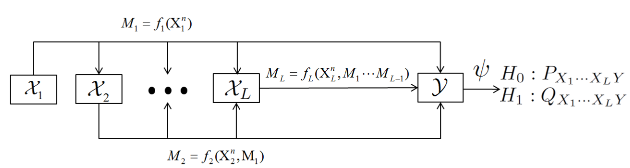

Fig. 1. Model

More specifically, terminal $\mathcal{X}_1$ uses an encoder

$$f_1 : \mathcal{X}_1^n \to \mathcal{M}_1 = \{1, 2, \ldots, M_1\}, \tag{2}$$

which is a map from $\mathcal{X}^n$ to $\mathcal{M}_1$. Terminal $\mathcal{X}_l,\ l=2,\cdots,L$ uses an encoder

$$f_l:(\mathcal{X}_l^n,\mathcal{M}_1,\cdots,\mathcal{M}_{l-1})\rightarrow\mathcal{M}_l=\{1,2,\dots,M_l\},\quad (3)$$

with rates R_l such that

$$\limsup_{n\rightarrow\infty}\frac{1}{n}\log M_l\leq R_l,\quad l=1,\cdots,L.\quad (4)$$

We also use the notation $||f_l||$ to denote the cardinality of f_l, $l=1,\cdots,L$. Hence, we have $||f_l||=M_l$, $l=1,\cdots,L$.

Using its own observations and messages received from encoders, terminal $\mathcal{Y}$ will use a decoding function ψ to decide which hypothesis is true:

$$\psi:(\mathcal{M}_1,\cdots,\mathcal{M}_L,\mathcal{Y}^n)\rightarrow\{H_0,H_1\}.\quad (5)$$

For any given $f_l,\ l=1,\cdots,L$ and ψ, one can define the acceptance region as

$$\begin{aligned}\mathcal{A}_n=\{&(x_1^n,\cdots,x_L^n,y^n)\in\mathcal{X}_1^n\times\cdots\times\mathcal{X}_L^n\times\mathcal{Y}^n:\\&\psi(f_1(x_1^n),\cdots,f_L(x_L^n),y^n)=H_0\}.\end{aligned}\quad (6)$$

Correspondingly, the type 1 error probability is defined as

$$\alpha_n=P^n_{X_1\cdots X_LY}(\bar{\mathcal{A}}_n),\quad (7)$$

in which $\bar{\mathcal{A}}_n$ denotes the complement set of $\mathcal{A}_n$. The type 2 error probability is defined as

$$\beta_n=Q^n_{X_1\cdots X_LY}(\mathcal{A}_n).\quad (8)$$

Our goal is to design the encoding functions $f_l,\ l=1,\cdots,L$ and the decoding function ψ to maximize the type 2 error exponent under the constraints that the type 1 error probability is less than ϵ and the communication rates satisfy (4).

More specifically, for a given $\epsilon>0$, we require

$$\alpha_n\leq\epsilon,\quad (9)$$

and define

$$\theta(R_1,\cdots,R_L,\epsilon)=\liminf_{n\rightarrow\infty}\left(-\frac{1}{n}\log\left(\min_{f_1,\cdots,f_L,\psi}\beta_n\right)\right),\quad (10)$$

in which the minimization is over all $f_1,\cdots,f_L,\psi$ satisfying conditions (4) and (9). With these notations, our goal mentioned above is then to characterize $\theta(R_1,\cdots,R_L,\epsilon)$, which is called the type 2 error exponent for the hypothesis testing H_0 versus H_1 with constraints (4) and (9).

B. Comparison with the Non-interactive Model

The main difference between our model and the non-interactive communication model considered in the existing works [14]–[16] is that, in the non-interactive communication model, the encoding function of each user relies only on its own observations. That is, instead of using (3) as in the cascaded case, the encoding function at terminal $\mathcal{X}_l$ in the non-interactive communication model is given as

$$f_l:\mathcal{X}_l^n\rightarrow\mathcal{M}_l=\{1,2,\dots,M_l\},\ l=1,\cdots,L.\quad (11)$$

All remaining definitions are the same. In the following, we will use $\theta_{\text{non-interactive}}(R_1,\cdots,R_L,\epsilon)$ to denote the corresponding type 2 error exponent in the non-interactive model.

C. Notation

Following [21], for any sequence $x^n=(x(1),\cdots,x(n))\in\mathcal{X}^n$, we use $n(a|x^n)$ to denote the total number of indices t at which $x(t)=a$. Then, the relative frequencies or empirical PMF- $\pi(a|x^n)\triangleq n(a|x^n)/n,\forall a\in\mathcal{X}$ of the components of x^n, is called the type of x^n and is denoted by $tp(x^n)$. The set of all types of sequences in $\mathcal{X}^n$ is denoted by $\mathcal{P}^n(\mathcal{X})$. Furthermore, we call a random variable $X^{(n)}$ that has the same distribution as $tp(x^n)$ as the type variable of x^n.

For a given a type $P_X\in\mathcal{P}^n(\mathcal{X})$ and a constant η, we denote by $T_\eta^{(n)}(X)$ the set of (P_X,η)-typical sequences in $\mathcal{X}^n$:

$$\begin{aligned}&T_\eta^{(n)}(X)\triangleq\\&\quad\{x^n\in\mathcal{X}^n:|\pi(a|x^n)-P_X(a)|\leq\eta P_X(a),\forall a\in\mathcal{X}\}.\end{aligned}\quad (12)$$

In the same manner, we use $\tilde{T}_\eta^{(n)}(X)$ to denote the set of $(\tilde{P}_X,\eta)$-typical sequences. Note that when $\eta=0$, $T_0^{(n)}(X)$ denote the set of sequences $x^n\in\mathcal{X}^n$ of type P_X, and we use $T^{(n)}(X)$ for simplicity.

Furthermore, for $y^n\in\mathcal{Y}^n$, we define $T_\eta^{(n)}(X|y^n)$ as the set of all x^ns that are jointly typical with y^n:

$$T_\eta^{(n)}(X|y^n)=\{x^n\in\mathcal{X}^n:(x^n,y^n)\in T_\eta^{(n)}(XY)\}.\quad (13)$$

III. Main Results

In this section, we focus on a special case: testing against independence, in which we are interested in determining whether $(X_1,\cdots,X_L)$ and Y are independent or not. In the case of testing against independence, $Q_{X_1\cdots X_LY}$ in (1) takes a special form

$$Q_{X_1\cdots X_LY}=P_{X_1\cdots X_L}P_Y$$

and two hypotheses in (1) become

$$H_0:P_{X_1\cdots X_LY}\quad\text{vs}\quad H_1:P_{X_1\cdots X_L}P_Y.\quad (14)$$

Note that the marginal distribution of $(X_1,\cdots,X_L)$ and Y are the same under both hypotheses in the case of testing against independence.

To simplify our presentation, we first present the results and detailed proof for $L=2$ case in Section III-A, and then extend the results of the $L=2$ case to the general case with $L\geq 2$ terminals in Section III-B.

A. $L=2$ Case

In this subsection, we study the $L=2$ case in detail. Our goal is to characterize the type 2 error exponent $\theta(R_1,R_2,\epsilon)$ under $\alpha_n\leq\epsilon$. We will show this in two parts. First, we design a scheme and characterize the corresponding error exponent for PMFs shown in (14). Then we will show that the scheme is optimal.

Compared with the non-interactive scenario considered in [14], in our model, $\mathcal{X}_2$ can use the message $f_1(X_1^n)$ from terminal $\mathcal{X}_1$ to perform the encoding. Hence, the coding scheme will be more complex while terminal $\mathcal{Y}$ could potentially receive more information. In the following, we first design a scheme and characterize its error exponent.

Theorem 1. *For the test against independence with $L = 2$ cascaded encoders, the best error exponent for the type 2 error probability satisfies*

$$\theta(R_1, R_2, \epsilon) \geq \max_{U_1U_2 \in \varphi_0} I(U_1U_2; Y) \tag{15}$$

where

$$\begin{aligned}\varphi_0 = \{U_1U_2 : R_1 \geq I(U_1; X_1),\ & R_2 \geq I(U_2; X_2|U_1), \\ & U_1 \leftrightarrow X_1 \leftrightarrow (X_2, Y), \qquad (16)\\ & U_2 \leftrightarrow (X_2, U_1) \leftrightarrow (X_1, Y), \qquad (17)\\ |\mathcal{U}_1| \leq |\mathcal{X}_1| + 1, & |\mathcal{U}_2| \leq |\mathcal{X}_2| \cdot |\mathcal{U}_1| + 1\}.\end{aligned}$$

Proof. In the following, $\eta > \eta' > \eta'' > \eta'''$ are given small numbers.

Codebook generation. Fix a joint distribution attaining the maximum in (15), which satisfies $P_{U_1U_2|X_1X_2Y} = P_{U_1|X_1}P_{U_2|U_1X_2}$. Let $P_{U_1}(u_1) = \sum_{x_1} P_{X_1}(x_1)P_{U_1|X_1}(u_1|x_1)$, and $P_{U_2|U_1}(u_2|u_1) = \sum_{x_2} P_{X_2|U_1}(x_2|u_1)P_{U_2|U_1X_2}(u_2|u_1, x_2)$. Randomly and independently generate $\lfloor 2^{nR_1} \rfloor$ sequences $u_1^n(m_1)$, $m_1 \in \{1, \cdots, \lfloor 2^{nR_1} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_1}(u_{1i})$. For each $u_1^n(m_1)$, randomly and independently generate $\lfloor 2^{nR_2} \rfloor$ sequences $u_2^n(m_2)$, $m_2 \in \{1, \cdots, \lfloor 2^{nR_2} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_2|U_1}(u_{2i}|u_{1i})$. These sequences constitute the codebook, which is revealed to all terminals. This process is shown in Fig. 2.

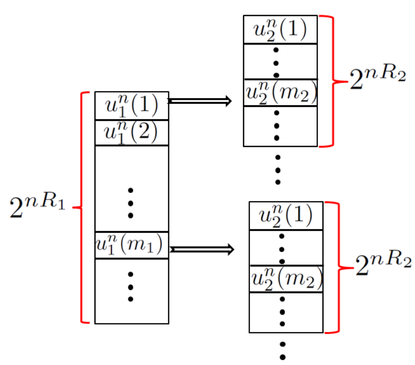

Fig. 2. Codebook generation

Encoding for terminal $\mathcal{X}_1$. Given a sequence x_1^n, terminal $\mathcal{X}_1$ finds a $u_1^n(m_1)$ such that $(x_1^n, u_1^n(m_1)) \in T_{\eta'''}^{(n)}(X_1U_1)$, then it sends the index m_1 to both terminal $\mathcal{X}_2$ and $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends 0.

Encoding for terminal $\mathcal{X}_2$. If $m_1 = 0$ is received from terminal $\mathcal{X}_1$, terminal $\mathcal{X}_2$ sends $m_2 = 0$ to terminal $\mathcal{Y}$. If $m_1 \neq 0$ is received, given x_2^n and m_1, terminal $\mathcal{X}_2$ finds a $u_2^n(m_2)$ such that $(u_1^n(m_1), u_2^n(m_2), x_2^n) \in T_{\eta''}^{(n)}(U_1U_2X_2)$ and sends the index m_2 to terminal $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends 0.

Testing. Upon receiving messages from terminal $\mathcal{X}_1$ and $\mathcal{X}_2$, terminal $\mathcal{Y}$ sets the acceptance region $\mathcal{A}_n$ for H_0 to

$$\mathcal{A}_n = \left\{(m_1, m_2, y^n) : (u_1^n(m_1), u_2^n(m_2), y^n) \in T_\eta^{(n)}(U_1U_2Y)\right\}.$$

This implies that terminal $\mathcal{Y}$ decides $\hat{H} = H_0$ if and only if no 0 is received and $(u_1^n(m_1), u_2^n(m_2), y^n) \in T_\eta^{(n)}(U_1U_2Y)$.

Analysis of two types of errors. Terminal $\mathcal{Y}$ chooses $\hat{H} = H_1$ if and only if one or more of the following events occur:

$$\begin{aligned}\varepsilon_1 &= \{(U_1^n(m_1), X_1^n) \notin T_{\eta'''}^{(n)}(U_1X_1) \\ &\qquad\qquad \text{for all } m_1 \in [1 : \lfloor 2^{nR_1} \rfloor]\},\\ \varepsilon_2 &= \{(U_1^n(M_1), U_2^n(m_2), X_2^n) \notin T_{\eta''}^{(n)}(U_1U_2X_2) \\ &\qquad\qquad \text{for all } m_2 \in [1 : \lfloor 2^{nR_2} \rfloor]\},\\ \varepsilon_3 &= \{(U_1^n(M_1), U_2^n(M_2), Y^n) \notin T_\eta^{(n)}(U_1U_2Y)\}.\end{aligned}$$

Here, we can see that $\bar{\mathcal{A}}_n = \varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3$.

a) Type 1 error probability: To compute the type 1 error probability, we assume that H_0 is true. Then

$$\begin{aligned}&\alpha_n = P_{X_1X_2Y}^n(\bar{\mathcal{A}}_n) = P_{X_1X_2Y}^n(\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3)\\ &\leq P_{X_1X_2Y}^n(\varepsilon_1) + P_{X_1X_2Y}^n(\bar{\varepsilon}_1 \cap \varepsilon_2) + P_{X_1X_2Y}^n(\bar{\varepsilon}_1 \cap \bar{\varepsilon}_2 \cap \varepsilon_3).\end{aligned}$$

We now bound each term.

1) By the covering lemma [21, Section 3.7], $P_{X_1X_2Y}^n(\varepsilon_1) \rightarrow 0$ as $n \rightarrow \infty$ if $R_1 \geq I(U_1; X_1) + \delta(\eta''')$.
2) Since $\eta'' > \eta'''$, $\bar{\varepsilon}_1 = \{(U_1^n(M_1), X_1^n) \in T_{\eta'''}^{(n)}(U_1X_1)\}$ and $X_2^n|\{X_1^n, U_1^n\} = X_2^n|X_1^n \sim \prod_{i=1}^n P_{X_2|X_1}(x_{2i}|x_{1i})$, by the conditional typicality lemma [21, Section 2.5], then $\Pr\{(U_1^n(M_1), X_1^n, X_2^n) \in T_{\eta''}^{(n)}(U_1X_1X_2)\} \rightarrow 1$, thus $\Pr\{(U_1^n(M_1), X_2^n) \in T_{\eta''}^{(n)}(U_1X_2)\} \rightarrow 1$ as $n \rightarrow \infty$. Therefore, again by the covering lemma, $P_{X_1X_2Y}^n(\bar{\varepsilon}_1 \cap \varepsilon_2) \rightarrow 0$ as $n \rightarrow \infty$ if $R_2 \geq I(U_2; X_2|U_1) + \delta(\eta'')$.
3) To bound the last term, we need two steps.
   Step 1: Since $X_2^n, Y^n|\{X_1^n = x_1^n, U_1^n(M_1) = u_1^n\} \sim \prod_{i=1}^n P_{X_2Y|X_1}(x_{2i}, y_i|x_{1i})$, we can show that $\Pr\{(X_1^n, X_2^n, U_1^n(M_1), Y^n) \in T_{\eta'}^{(n)}(X_1X_2U_1Y)\} \rightarrow 1$ using the conditional typicality lemma.
   Step 2: Since we have the Markov chain $U_2^n(M_2) \leftrightarrow (X_2^n, U_1^n(M_1)) \leftrightarrow (X_1^n, Y^n)$ and $(X_1^n, X_2^n, U_1^n(M_1), Y^n) \in T_{\eta'}^{(n)}(X_1X_2U_1Y)$ by *Step 1*, we can show that $\Pr\{(U_1^n(M_1), U_2^n(M_2), X_1^n, X_2^n, Y^n) \in T_\eta^{(n)}(U_1U_2X_1X_2Y)\} \rightarrow 1$ as $n \rightarrow \infty$ using Markov lemma [21, Section 12.1].

b) Type 2 error probability: To calculate the type 2 error probability, assume in this case that H_1 is true, then we have

$$\begin{aligned}\beta_n &= (P_{X_1X_2}P_Y)^n(\mathcal{A}_n) = (P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_1 \cap \bar{\varepsilon}_2 \cap \bar{\varepsilon}_3)\\ &= (P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_1) \cdot (P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_2|\bar{\varepsilon}_1)\\ &\qquad \cdot (P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_3|\bar{\varepsilon}_1 \cap \bar{\varepsilon}_2).\end{aligned}$$

We now bound each factor.

1) By the covering lemma, $(P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_1) \to 1$ as $n \to \infty$, if $R_1 \geq I(U_1; X_1) + \delta(\eta''')$.
2) The second term is the same as that of H_0 as it depends only on $P^n_{X_1X_2}$. Hence again by the covering lemma, $(P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_2|\bar{\varepsilon}_1) \to 1$ as $n \to \infty$ if $R_2 \geq I(U_2; X_2|U_1) + \delta(\eta'')$.
3) For the third term, we have

$$
\begin{aligned}
&(P_{X_1X_2}P_Y)^n(\bar{\varepsilon}_3|\bar{\varepsilon}_1 \cap \bar{\varepsilon}_2) \\
&= \sum_{(u_1^n,u_2^n,y^n)\in T_\eta^{(n)}(U_1U_2Y)} (P_{X_1X_2}P_Y)^n\{U_1^n(M_1) = u_1^n, \\
&\qquad U_2^n(M_2) = u_2^n, Y^n = y^n|\bar{\varepsilon}_1 \cap \bar{\varepsilon}_2\} \\
&\leq 2^{n(H(U_1U_2Y)+\delta(\eta))}2^{-n(H(U_1U_2)-\delta(\eta'))}2^{-n(H(Y)-\delta(\eta'))} \\
&= 2^{-n(I(U_1U_2;Y)-\delta(\eta))}.
\end{aligned}
$$

Combining the bounds on these three factors, we have

$$\beta_n \leq 2^{-n(I(U_1U_2;Y)-\delta(\eta))}.$$

This completes the achievability proof. □

Now we show that the scheme in Theorem 1 is optimal.

Theorem 2. *In the testing against independence with $L = 2$ cascaded encoders, when the type 1 error constraint (9) is satisfied, the best error exponent for the type 2 error probability satisfies*

$$\lim_{\epsilon\to 0} \theta(R_1, R_2, \epsilon) \leq \max_{U_1U_2\in\varphi_0} I(U_1U_2; Y) \tag{18}$$

where φ_0 is defined in Theorem 1.

Proof. First, for any scheme (f_1, f_2, ψ) that satisfies the type 1 error constraint (9) and rate constraints (4), we have

$$
\begin{aligned}
&D(P_{M_1M_2Y^n}||P_{M_1M_2}P_{Y^n}) \\
&\overset{(a)}{\geq} (1-\alpha_n)\log\frac{1-\alpha_n}{\beta_n} + \alpha_n \log\frac{\alpha_n}{1-\beta_n} \\
&\overset{(b)}{\geq} (1-\epsilon)\log\frac{1}{\beta_n} - H(\alpha_n),
\end{aligned}
$$

in which $M_1 = f_1(X_1^n)$, $M_2 = f_2(X_2^n, M_1)$, α_n and β_n are defined in (7) and (8) respectively, and $H(\alpha_n) := -(1-\alpha_n)\log(1-\alpha_n) - \alpha_n \log \alpha_n$ is the binary entropy function. In the above derivation, (a) is true due to the log sum inequality [21], and (b) follows by the constraint (9). By the communication constraints (4), we have $H(M_l) \leq nR_l,\ l = 1, 2$.

Hence we have the following multi-letter expression of the upper bound

$$
\begin{aligned}
\lim_{\epsilon\to 0} \theta(R_1, R_2, \epsilon) &\leq \lim_{n\to\infty} \frac{1}{n} D(P_{M_1M_2Y^n}||P_{M_1M_2}P_{Y^n}) \\
&= \lim_{n\to\infty} \frac{1}{n}(I(M_1M_2; Y^n)) \\
&= H(Y) - \lim_{n\to\infty} \frac{1}{n}(H(Y^n|M_1M_2)). \quad (19)
\end{aligned}
$$

Then, we single-letterize the upper bound in (19) in the following steps. First consider

$$
\begin{aligned}
nR_1 &\geq H(M_1) \geq I(M_1; X_1^nX_2^n) \\
&= \sum_{i=1}^n I(M_1X_1^{i-1}X_{2(i+1)}^n; X_{1i}X_{2i}) \\
&\geq \sum_{i=1}^n I(M_1X_1^{i-1}X_{2(i+1)}^n; X_{1i}) \\
&\overset{(a)}{=} \sum_{i=1}^n I(U_{1i}; X_{1i}),
\end{aligned}
$$

where (a) is true by identifying $U_{1i} = (M_1, X_1^{i-1}, X_{2(i+1)}^n)$ and noting that $U_{1i} \leftrightarrow X_{1i} \leftrightarrow (X_{2i}, Y_i)$ forms a Markov chain.

Next consider

$$
\begin{aligned}
nR_2 &\geq H(M_2) \geq I(M_2; X_1^nX_2^nY^n|M_1) \\
&= \sum_{i=1}^n I(M_2; X_{1i}X_{2i}Y_i|M_1X_1^{i-1}X_{2(i+1)}^nY^{i-1}) \\
&\overset{(b)}{=} \sum_{i=1}^n I(M_2Y^{i-1}; X_{1i}X_{2i}Y_i|M_1X_1^{i-1}X_{2(i+1)}^n) \\
&\geq \sum_{i=1}^n I(M_2Y^{i-1}; X_{2i}|M_1X_1^{i-1}X_{2(i+1)}^n) \\
&\overset{(c)}{=} \sum_{i=1}^n I(U_{2i}; X_{2i}|U_{1i}),
\end{aligned}
$$

where (b) is true since $Y^{i-1} \leftrightarrow (X_{2(i+1)}^n, X_1^{i-1}, M_1) \leftrightarrow (X_{1i}, X_{2i}, Y_i)$ forms a Markov chain, which can be derived in the following way,

$$
\begin{aligned}
&(X_1^n, X_{1i}, X_{2i}, Y_i, X_{2(i+1)}^n) \leftrightarrow X_1^{i-1} \leftrightarrow Y^{i-1} \\
&\Rightarrow (M_1, X_{1i}, X_{2i}, Y_i, X_{2(i+1)}^n) \leftrightarrow X_1^{i-1} \leftrightarrow Y^{i-1} \\
&\overset{(d)}{\Rightarrow} (X_{1i}, X_{2i}, Y_i) \leftrightarrow (M_1, X_1^{i-1}, X_{2(i+1)}^n) \leftrightarrow Y^{i-1},
\end{aligned}
$$

in which (d) follows by the weak union property of Markov chain [22]. (c) follows by defining $U_{2i} = (M_2, Y^{i-1})$ and noting that $U_{2i} \leftrightarrow (U_{1i}, X_{2i}) \leftrightarrow (X_{1i}, Y_i)$ forms a Markov chain which is proved in Appendix A.

Finally, we consider

$$
\begin{aligned}
H(Y^n|M_1M_2) &= \sum_{i=1}^n H(Y_i|M_1M_2Y^{i-1}) \\
&\geq \sum_{i=1}^n H(Y_i|M_1M_2Y^{i-1}X_1^{i-1}X_{2(i+1)}^n) \\
&= \sum_{i=1}^n H(Y_i|U_{1i}U_{2i}).
\end{aligned}
$$

Define the time-sharing random variable $Q \sim \text{Unif}[1:n]$ and independent of $(M_1, M_2, X_1^n, X_2^n, Y^n)$, and identify $U_1 = (U_{1Q}, Q)$, $U_2 = (U_{2Q}, Q)$, $X_1 = X_{1Q}$, $X_2 = X_{2Q}$, and $Y = Y_Q$. Clearly, we have $U_1 \leftrightarrow X_1 \leftrightarrow (X_2, Y)$ and $U_2 \leftrightarrow (U_1, X_2) \leftrightarrow (X_1, Y)$ form two Markov chains. Hence we have shown

$$R_1 \geq I(U_1;X_1), \quad R_2 \geq I(U_2;X_2|U_1),$$
$$\lim_{\epsilon\to 0}\theta(R_1,R_2,\epsilon) \leq H(Y)-H(Y|U_1U_2)=I(Y;U_1U_2),$$

for $P_{U_1U_2|X_1X_2Y} = P_{U_1|X_1}P_{U_2|U_1X_2}$. This completes the converse proof. □

Hence, we obtain a matching upper and lower bound on the type 2 error exponent which is shown in Theorem 3.

Theorem 3. *In the testing against independence with $L=2$ cascaded encoders, when the type 1 error constraint* (9) *is satisfied, the best error exponent for the type 2 error probability satisfies*

$$\lim_{\epsilon\to 0}\theta(R_1,R_2,\epsilon) = \max_{U_1U_2\in\varphi_0} I(U_1U_2;Y) \tag{20}$$

where φ_0 is defined in Theorem 1.

B. General L Case

The results in the previous subsection can be extended to the general case with L terminals and a decision maker $\mathcal{Y}$. The result is shown in the following theorem. The proof of this theorem can be obtained by properly modifying the proof for $L=2$ case, and hence is omitted for brevity.

Theorem 4. *In the testing against independence with L cascaded encoders, the best type 2 error exponent satisfies*

$$\theta(R_1,\cdots,R_L,\epsilon) \geq \max_{U_1\cdots U_L\in\varphi} I(U_1\cdots U_L;Y), \tag{21}$$

in which

$$\begin{aligned}\varphi = \{&U_1\cdots U_L : R_1 \geq I(U_1;X_1),\\ &R_l \geq I(U_l;X_l|U_1\cdots U_{l-1}),\\ &U_1 \leftrightarrow X_1 \leftrightarrow (X_2,\cdots,X_L,Y), \qquad (22)\\ &U_l \leftrightarrow (X_l,U_1,\cdots,U_{l-1}) \leftrightarrow (X_1,\cdots,X_{l-1},\\ &\qquad X_{l+1},\cdots,X_L,Y), \qquad (23)\\ &|\mathcal{U}_1| \leq |\mathcal{X}_1|+1,\\ &|\mathcal{U}_l| \leq |\mathcal{X}_l|\cdot|\mathcal{U}_{l-1}|\cdots|\mathcal{U}_1|+1, l=2,\cdots,L\}.\end{aligned}$$

C. General PMF Case

Furthermore, we extend our study to the general PMF case (i.e., not necessarily for the test against independence anymore) with L terminals:

$$H_0 : P_{X_1\cdots X_LY}, \quad H_1 : Q_{X_1\cdots X_LY}.$$

The result is shown in the following theorem.

Theorem 5. *For the case with general hypothesis $P_{X_1\cdots X_LY}$ vs $Q_{X_1\cdots X_LY}$ with L cascaded encoders, the best error exponent of the type 2 error probability satisfies*

$$\theta(R_1,\cdots,R_2,\epsilon) \geq \max_{U_1\cdots U_L\in\varphi} \tag{24}$$
$$\min_{\tilde{P}_{U_1\cdots U_LX_1\cdots X_LY}\in\xi} D(\tilde{P}_{U_1\cdots U_LX_1\cdots X_LY}||Q_{U_1\cdots U_LX_1\cdots X_LY}),$$

where φ is defined in Theorem 4,

$$\xi = \{\tilde{P}_{U_1\cdots U_LX_1\cdots X_LY} : \tilde{P}_{U_1\cdots U_lX_l} = P_{U_1\cdots U_lX_l},$$
$$\tilde{P}_{U_1\cdots U_LY} = P_{U_1\cdots U_LY},\ l=1,\cdots,L\},$$

$Q_{U1|X_1} = P_{U_1|X_1}$ *and* $Q_{U_l|U_1\cdots U_{l-1}X_l} = P_{U_l|U_1\cdots U_{l-1}X_l}$ *for* $l=2,\cdots,L$.

Proof. We can adopt the coding scheme in Section III-A and the analysis in [11] with necessary changes. Details are omitted for brevity. □

IV. Comparison with the Non-interactive Communication Model

In this section, we compare the performance achieved by the cascaded communication scheme and that of the non-interactive communication scheme. We will provide concrete examples to show that for certain PMF and positive communication rates, the scheme with cascaded communication outperforms that of the non-interactive communication scheme. On the other hand, we will also prove that when the communication rates go to zero (zero-rate compression), the performance of the cascaded communication scheme is the same as that of the non-interactive communication scheme, and hence the cascaded scheme does not improve the performance in these scenarios.

A. Example When the Cascaded Scheme Is Better Than the Non-interactive Scheme

Here, we provide an example in which the error exponent achieved using the cascaded scheme is larger than that can be achieved using the non-interactive scheme. The example is about the testing against independence case. The testing against independence problem with non-interactive communications and multiple terminals was studied in [14], which provides a lower and an upper bound on the type 2 error exponent of non-interactive schemes. As the lower and upper bounds in [14] do not match with each other, in this part, we compare the type 2 error exponent achieved by the scheme with cascaded communications shown in the proof of Theorem 1, with the upper bound on the type 2 error exponent of the non-interactive scheme derived in Theorem IV3 of [14].

In the example, we let $X_1,\ X_2$ and Y be binary random variables with joint PMF $P_{X_1X_2Y}$.

TABLE I
The joint PMF $P_{X_1X_2Y}$

X_1X_2Y	000	010	100	110
$P_{X_1X_2Y}$	0.0704	0.2108	0.0015	0.3233
X_1X_2Y	001	011	101	111
$P_{X_1X_2Y}$	0.2206	0.0667	0.0046	0.1021

For testing against independence case, we have $Q_{X_1X_2Y} = P_{X_1X_2}P_Y$, which can be easily calculated from Table I. With given communication constraint $R=R_1=R_2$, we use Theorem 1 to find the best value of the type 2 error exponent that we can achieve using our cascaded scheme. For comparison, we also use Theorem IV3 of [14] to find an upper bound on the type 2 error exponent of the non-interactive case. For $R = 0.48$, we list the conditional distributions $P_{U_1|X_1}$ and $P_{U_2|X_2}$ for non-interactive case and the conditional distributions $P_{U_1|X_1}$ and $P_{U_2|X_2U_1}$ for the cascaded case in the following tables.

TABLE II
$P_{U_1|X_1}$ AND $P_{U_2|X_2}$ FOR NON-INTERACTIVE CASE WHEN $R = 0.48$

$U_1\|X_1$	$0\|0$	$1\|0$	$0\|1$	$1\|1$
$P_{U_1\|X_1}$	0.9991	0.0009	0.1564	0.8436
$U_2\|X_2$	$0\|0$	$1\|0$	$0\|1$	$1\|1$
$P_{U_2\|X_2}$	0.9686	0.0314	0.0357	0.9643

TABLE III
$P_{U_1|X_1}$ AND $P_{U_2|X_2U_1}$ FOR CASCADED CASE WHEN $R = 0.48$

$U_1\|X_1$	$0\|0$	$1\|0$	$0\|1$	$1\|1$
$P_{U_1\|X_1}$	0.0155	0.9845	0.5829	0.4171
$U_2\|X_2U_1$	$0\|00$	$1\|00$	$0\|01$	$1\|01$
$P_{U_2\|X_2U_1}$	0.0636	0.9364	0.9727	0.0273
$U_2\|X_2U_1$	$0\|10$	$1\|10$	$0\|11$	$1\|11$
$P_{U_2\|X_2U_1}$	0.9898	0.0102	0.0005	0.9995

The simulation results for different Rs are shown in Fig. 3. From Fig. 3, we can see that the type 2 error exponents in both cases increase with the increasing value of R, which makes sense as the more information we can send, the less errors we will make. We also observe that the type 2 error exponent achieved using our cascaded communication scheme is even larger than an upper bound on the type 2 error exponent of any non-interactive schemes. Hence, we confirm the intuitive idea that more information offered by the cascaded communication facilitates a better decision making for certain testing against independence cases with positive communication rates.

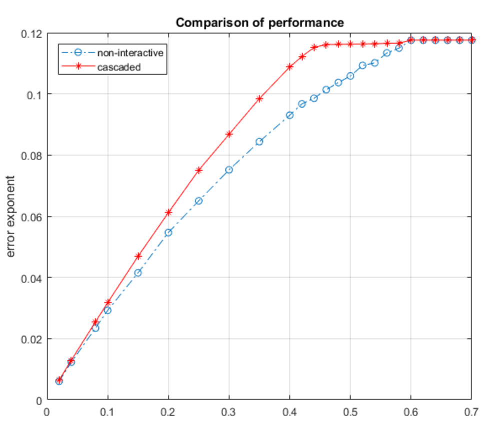

Fig. 3. Simulation results

We also list the error exponents for $R \geq 0.46$ in TABLE IV since it is not obvious to see the increase in the performance of both cascaded and non-interactive communication. Note that when R is large enough, we can let $U_1 = X_1$ and $U_2 = X_2$. And we have $\theta_{\text{non-ineractive}} \leq I(X_1X_2;Y) = \theta$. To achieve this maximum value, the constraints of R_1 and R_2 can be simplified in the following:

- Non-interactive case:

$$R_1 \geq H(X_1), \quad R_2 \geq H(X_2). \tag{25}$$

- Cascaded case:

$$R_1 \geq H(X_1), \quad R_2 \geq H(X_2|X_1). \tag{26}$$

We also list the values these theoretic limits in TABLE V.

TABLE IV
ERROR EXPONENTS FOR $R \geq 0.42$

R	0.42	0.46	0.50	0.52
$\theta_{\text{non-interactive}}$	0.096724	0.10136	0.10585	0.10930
θ	0.11222	0.11612	0.11625	0.11630
R	0.54	0.58	0.62	0.66
$\theta_{\text{non-interactive}}$	0.11013	0.11496	0.11754	0.11754
θ	0.11640	0.11661	0.11755	0.11755
R	0.68	0.70	0.72	0.74
$\theta_{\text{non-interactive}}$	0.11754	0.11754	0.11754	0.11754
θ	0.11755	0.11755	0.11755	0.11755

TABLE V
THEORETIC LIMITS FOR $U_1 = X_1$ AND $U_1 = X_2$

$I(X_1X_2;Y)$	$H(X_1)$	$H(X_2)$	$H(X_2\|X_1)$
0.1187	0.6838	0.5127	0.4259

From TABLE IV and TABLE V, we can see that the increasing speed of θ decreases when $R \geq 0.42$ as R is large enough for terminal $\mathcal{X}_2$. This same happens for $\theta_{\text{non-interactive}}$ when $R \geq 0.52$. Furthermore, both $\theta_{\text{non-interactive}}$ and θ approach the best possible value of $I(X_1, X_2; Y) = 0.1187$ as $R \to 0.68$. We note that there is a slight gap between the theoretic limit and simulation results. This is due to the precision of the numerical simulation.

B. Example When the Cascaded Scheme Has the Same Performance as that of the Non-interactive Scheme

In this subsection, we provide an example for which the cascaded scheme has the same performance as that of the non-interactive scheme. In particular, we will prove that, under "zero-rate" data compression, i.e. $R_l = +0, \ l = 1, \cdots, L$, cascaded communication does not improve the performance. More specifically, we consider our model testing the general hypotheses under the communication constraints

$$\text{as } n \to \infty, \ M_l \to \infty, \tag{27}$$

but

$$R_l = \frac{1}{n} \log M_l \to 0, \ l = 1, 2. \tag{28}$$

In the non-interactive communication scenario with zero-rate compression, a matching upper bound and lower bound on the type 2 error exponent was provided in [13, Theorem 2] when $Q_{X_1\cdots X_L Y} > 0$. If we can prove an upper bound on the type 2 error exponent for the cascaded communication case that is no larger the error exponent shown in [13], then we can arrive at the conclusion that the cascaded communication won't help under the zero-rate compression case.

For reference, we state the error exponent of the non-interactive scheme characterized in [13] in the following.

Theorem 6. *([13]) Let $P_{X_1\cdots X_L Y}$ be arbitrary and $Q_{X_1\cdots X_L Y} > 0$, for all $\epsilon \in [0,1)$, the type 2 error exponent for zero-rate compression under $\alpha_n \leq \epsilon$ with L non-interactive encoders is given by*

$$\begin{aligned}&\theta_{\textit{non-interactive}}(+0,\cdots,+0,\epsilon)\\&\quad= \min_{\tilde{P}_{X_1\cdots X_L Y}\in\mathcal{L}} D(\tilde{P}_{X_1\cdots X_L Y}||Q_{X_1\cdots X_L Y}) \qquad (29)\end{aligned}$$

where

$$\mathcal{L} = \Big\{\tilde{P}_{X_1\cdots X_L Y} : \tilde{P}_{X_l} = P_{X_l},\ l = 1,\cdots,L,\ \tilde{P}_Y = P_Y\Big\}.$$

In the following, we provide an upper bound on the type 2 error exponent for the cascaded case.

Theorem 7. *Let $P_{X_1\cdots X_L Y}$ be arbitrary and $Q_{X_1\cdots X_L Y} > 0$, for all $\epsilon \in [0,1)$, the best type 2 error exponent for zero-rate compression under $\alpha_n \leq \epsilon$ with L cascaded encoders satisfies*

$$\begin{aligned}&\theta(+0,\cdots,+0,\epsilon)\\&\quad\leq \min_{\tilde{P}_{X_1\cdots X_L Y}\in\mathcal{L}} D(\tilde{P}_{X_1\cdots X_L Y}||Q_{X_1\cdots X_L Y}) \qquad (30)\end{aligned}$$

where $\mathcal{L}$ is defined in Theorem 6.

Proof. Please see Appendix B. □

Comparing Theorem 7 with Theorem 6, we can see that the upper bound on the type 2 error exponent for the cascaded communication scheme is the same as the type 2 error exponent achievable by the non-interactive communication scheme. This implies that the performance of the cascaded communication scheme is the same as that of the non-interactive communication scheme in the zero-rate data compression case.

The conclusion that cascaded communication does not improve the type 2 error exponent under the zero-rate data compression case also holds when we have a stronger constraint on the type 1 error probability:

$$\alpha_n \leq \exp[-nr],\ r > 0. \qquad (31)$$

This constraint is called exponential-type constraint.

In the cascaded communication case, based on the results in Theorem 7, we can use a similar strategy as in [10] to convert the problem under the exponential-type constraint (31) to the corresponding problem under the constraint in (9). As the converting strategy is independent of the communication style, it will be the same as that in [16]. Then an upper bound on the type 2 error exponent under the exponential-type constraint can be easily derived without going into details, shown in the sequel. Here, we use $\sigma(+0,\cdots,+0,r)$ to denote the type 2 error exponent under exponential-type constraint (31) for general case with L terminals.

Theorem 8. *Let $P_{X_1\cdots X_L Y}$ be arbitrary and $Q_{X_1\cdots X_L Y} > 0$, the best type 2 error exponent for zero-rate compression case under $\alpha_n \leq \exp[-nr]$ with L cascaded encoders satisfies*

$$\begin{aligned}&\sigma(+0,\cdots,+0,r)\\&\quad\leq \min_{\tilde{P}_{X_1\cdots X_L Y}\in\mathcal{H}_r} D(\tilde{P}_{X_1\cdots X_L Y}||Q_{X_1\cdots X_L Y}) \qquad (32)\end{aligned}$$

where

$$\begin{aligned}\mathcal{H}_r = \Big\{&\tilde{P}_{X_1\cdots X_L Y} : \tilde{P}_{X_l} = \hat{P}_{X_l}, \tilde{P}_Y = \hat{P}_Y, \quad l = 1,\cdots,L\\&\text{for some } \hat{P}_{X_1\cdots X_L Y} \in \varphi_r\Big\}, \qquad (33)\end{aligned}$$

$$\varphi_r = \{\hat{P}_{X_1\cdots X_L Y} : D(\hat{P}_{X_1\cdots X_L Y}||P_{X_1\cdots X_L Y}) \leq r\}. \qquad (34)$$

Comparing Theorem 8 with [16, Theorem 4], where a matching upper and lower bound is provided for the non-interactive scheme, we can conclude that there is no gain in performance on the type 2 error exponent under zero-rate compression with the exponential-type constraint on the type 1 error probability.

V. Conclusion

In the paper, we have considered distributed testing problems with cascaded encoders. We have first investigated the special case of testing against independence. We have designed a scheme to benefit from the extra information provided by cascaded communications, and have shown that the proposed scheme is optimal when certain Markovian relation exists. We have then derived a lower bound on the type 2 error exponent for cases with general hypotheses. Compared with existing results in the non-interactive communication cases, we have shown that cascaded communication does provide performance gain under certain PMFs and positive communication rates but does not offer gain under zero-rate data compression scenarios.

Appendix A
Proof of the Markov chain $U_{2i} \leftrightarrow (U_{1i}, X_{2i}) \leftrightarrow (X_{1i}, Y_i)$

First, we need the following lemma introduced and proved in [23, Lemma 1].

Lemma 1. *[23] Let A_1, A_2, B_1, B_2 be the random variables with joint PMF $P_{A_1A_2B_1B_2} = P_{A_1B_1}P_{A_2B_2}$ and assume that $\{f^i\}_{i=1}^k, \{g^i\}_{i=1}^k$ are any collection of P-measurable mappings with domain structure given by:*

$$\begin{aligned}&f^1(A_1,A_2); f^2(A_1,A_2,g^1);\cdots; f^k(A_1,A_2,g^1,\cdots,g^{k-1}),\\&g^1(B_1,B_2,f^1);\cdots; g^k(B_1,B_2,f^1,\cdots,f^k). \qquad (35)\end{aligned}$$

Then,

$$I(A_2;B_1|A_1,B_2,f^1,f^2,\cdots,f^k,g^1,g^2,\cdots,g^k) = 0. \qquad (36)$$

To prove the Markov chain $U_{2i} \leftrightarrow (U_{1i}, X_{2i}) \leftrightarrow (X_{1i}, Y_i)$, first we set

$$\begin{cases} A_1 := X_1^{i-1}, & B_1 := X_2^{i-1} \\ A_2 := X_{1i}^n, & B_2 := X_{2i}^n \end{cases} \tag{37}$$

Then according to Lemma 1, we have

$$I(X_{1i}^n; X_2^{i-1}|X_1^{i-1}, X_{2i}^n, M_1) = 0, \tag{38}$$

where $M_1 = f^1(A_1, A_2)$. Thus, we have the following Markov chain,

$$(X_{1i}, X_{1(i+1)}^n) \leftrightarrow (X_1^{i-1}, X_{2i}^n, M_1) \leftrightarrow X_2^{i-1}. \tag{39}$$

As $M_2 = g^1(B_1, B_2, M_1)$, we have

$$X_{1i} \leftrightarrow (X_1^{i-1}, X_{2i}^n, M_1) \leftrightarrow M_2. \tag{40}$$

Since $Y_i \leftrightarrow (X_{1i}, X_{2i}) \leftrightarrow (X_1^{i-1}, X_{2(i+1)}^n, M_1, M_2)$, we can have

$$(X_{1i}, Y_i) \leftrightarrow (X_1^{i-1}, X_{2i}, X_{2(i+1)}^n, M_1) \leftrightarrow (M_2, Y^{i-1}), \tag{41}$$

i.e.

$$(X_{1i}, Y_i) \leftrightarrow (X_{2i}, U_{1i}) \leftrightarrow U_{2i}. \tag{42}$$

Appendix B
Proof of Theorem 7

In this appendix, to facilitate the presentation, we show a detailed proof for $L = 2$. The proof for the general L is similar. Our proof follows a similar strategy as that in [13] and employs the "blowing-up" lemma [24].

First, we define

$$C_{m_1} = \{x_1^n \in \mathcal{X}_1^n : f_1(x_1^n) = m_1\},$$
$$D_{m_2|m_1} = \{x_2^n \in \mathcal{X}_2^n : f_2(x_2^n, m_1) = m_2\},$$
$$F_{m_1,m_2} = \{y^n \in \mathcal{Y}^n : \psi(m_1, m_2, y^n) = H_0\},$$

then we can write

$$\mathcal{A}_n = \bigcup_{m_1=1}^{||f_1||} \bigcup_{m_2=1}^{||f_2||} C_{m_1} \times D_{m_2|m_1} \times F_{m_1,m_2}. \tag{43}$$

And we can see that C_{m_1}s are pairwise disjoint and for fixed m_1, $D_{m_2|m_1}$s are pairwise disjoint for different m_2.

We have $P^n_{X_1X_2Y}(\mathcal{A}_n) \geq 1 - \epsilon$, then there exists an index (m_{10}, m_{20}) such that

$$P^n_{X_1X_2Y}(C_{m_{10}} \times D_{m_{20}|m_{10}} \times F_{m_{10},m_{20}}) \geq \frac{1-\epsilon}{||f_1|| \cdot ||f_2||}.$$

To simplify the notations, we let $C = C_{m_{10}}$, $D_{m_{20}|m_{10}} = D$ and $F_{m_{10},m_{20}} = F$. We can rewrite the equation above as

$$P^n_{X_1X_2Y}(C \times D \times F) \geq \exp(-n\delta_n) \tag{44}$$

where $\delta_n = -\frac{1}{n}\log(1-\epsilon) + \frac{1}{n}\log(||f_1|| \cdot ||f_2||)$ and $\delta_n \to 0$ by (27) and (28). (44) implies that

$$P^n_{X_1}(C) \geq \exp(-n\delta_n),\ P^n_{X_2}(D) \geq \exp(-n\delta_n),$$
$$P^n_Y(F) \geq \exp(-n\delta_n).$$

Define the Hamming k-neighborhood $\Gamma^k C$ of C by

$$\Gamma^k C = \{z^n \in \mathcal{X}_1^n : \exists x_1^n \in C,\ \text{s.t. } d(x_1^n, u^n) \leq k\}.$$

Using Blowing-up lemma [25], there exists sequences k_n and γ_n satisfying $k_n/n \to 0$ and $\gamma_n \to 0$, and such that

$$P^n_{X_1}(\Gamma^{k_n} C) \geq 1 - \gamma_n, \tag{45}$$
$$P^n_{X_2}(\Gamma^{k_n} D) \geq 1 - \gamma_n, \tag{46}$$
$$P^n_Y(\Gamma^{k_n} F) \geq 1 - \gamma_n. \tag{47}$$

Furthermore, k_n and γ_n depend only on $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{Y}$ and γ_n. In the following, we will use k instead of k_n. (45), (46) and (47) hold true if we replace P by $\tilde{P}$ where $\tilde{P}_{X_1X_2Y}$ satisfies the marginal constraints $\tilde{P}_{X_1} = P_{X_1}, \tilde{P}_{X_2} = P_{X_2}$, and $\tilde{P}_Y = P_Y$. Moreover, via simple derivations we have

$$\tilde{P}^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \geq 1 - 3\gamma_n. \tag{48}$$

As $\tilde{T}^{(n)}_\eta(X_1X_2Y)$ is the set of $(\tilde{P}_{X_1X_2Y}, \eta)$-typical sequences, then $\tilde{P}^n_{X_1X_2Y}(\tilde{T}^{(n)}_\eta(X_1X_2Y)) \geq 1 - \eta_n$, where η_n is a small number such that $\eta_n/n \to 0$ as $n \to \infty$. Hence, for all sufficiently large n, we obtain

$$\tilde{P}^n_{X_1X_2Y}\big((\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \tilde{T}^{(n)}_\eta(X_1X_2Y)\big) \geq \frac{1}{2}. \tag{49}$$

By the definition of $\tilde{T}^{(n)}_\eta(X_1X_2Y)$, we have the following decomposition:

$$\tilde{T}^{(n)}_\eta(X_1X_2Y) = \bigcup_{\substack{\hat{P}_{X_1X_2Y} \in \mathcal{P}^n(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}) \\ |\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y}}} \hat{T}^{(n)}(X_1X_2Y).$$

Given the fact of equiprobable elements of a given $\hat{T}^{(n)}(X_1X_2Y)$, (49) can be rewritten as

$$\sum_{\substack{\hat{P}_{X_1X_2Y} \in \mathcal{P}^n(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}) \\ |\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y}}} \tilde{P}^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y)) \frac{|(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)|}{|\hat{T}^{(n)}(X_1X_2Y)|} \geq \frac{1}{2}.$$

Hence, there exists a type $\hat{P}_{X_1X_2Y} \in \mathcal{P}^n(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y})$ satisfying $|\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y}$ and such that

$$\frac{|(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)|}{|\hat{T}^{(n)}(X_1X_2Y)|} \geq \frac{1}{2}.$$

Since pairs (x_1^n, x_2^n, y^n) of the same type are also equiprobable under $Q^n_{X_1X_2Y}$, we conclude that for the previous type $\hat{P}_{X_1X_2Y}$,

$$\begin{aligned} &Q^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \\ &\geq Q^n_{X_1X_2Y}((\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)) \\ &\geq \frac{1}{2} Q^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y)). \end{aligned} \tag{50}$$

Consider an arbitrary element (z^n, v^n, w^n) of $\Gamma^k C \times \Gamma^k D \times \Gamma^k F$. By definition of Γ^k, there exists at least one element $(x_1^n, x_2^n, y^n) \in C \times D \times F$ such that (x_{1i}, x_{2i}, y_i) differs from

(z_i, v_i, w_i) for at most $3k$ values of i. We thus have

$$Q^n_{X_1X_2Y}(z^n, v^n, w^n) = \prod_{i=1}^{n} Q_{X_1X_2Y}(z_i, v_i, w_i) = \rho^{-3k} Q_{X_1X_2Y}(x_1^n, x_2^n, y^n), \quad (51)$$

where $\rho = \min_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, y \in \mathcal{Y}} Q_{X_1X_2Y}(x_1, x_2, y) > 0$. As (z^n, v^n, w^n) ranges over $\Gamma^k C \times \Gamma^k D \times \Gamma^k F$, each element (x_1^n, x_2^n, y^n) of $C \times D \times F$ will be selected at most $|\Gamma^k(x_1^n)| \cdot |\Gamma^k(x_2^n)| \cdot |\Gamma^k(y^n)|$ times. By virtue of this, (51) yields

$$Q^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \leq \rho^{-3k} |\Gamma^k C(x_1^n)| \cdot |\Gamma^k(x_2^n)| \cdot |\Gamma^k(y^n)| Q^n_{X_1X_2Y}(C \times D \times F).$$

From [25], we have the upper bound

$$|\Gamma^k(x_1^n)| \leq \exp\left[n\left(H\left(\frac{k}{n}\right) + \frac{k}{n}\log|\mathcal{X}_1|\right)\right].$$

Thus, we can write

$$Q^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \leq \exp(n\xi_n) Q^n_{X_1X_2Y}(C \times D \times F), \quad (52)$$

where $\xi_n \to 0$.

Finally, combining (50) and (52) with the upper bound on $Q^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y))$, we have

$$Q^n_{X_1X_2Y}(C \times D \times F) \geq \exp[-n(D(\hat{P}_{X_1X_2Y}||Q_{X_1X_2Y}) + \varsigma_n)],$$

where $\varsigma_n \to 0$. Since $D(\tilde{P}_{X_1X_2Y}||\tilde{Q}_{X_1X_2Y})$ is uniformly continuous, we can find a sequence $\mu_n = \mu_n(\rho, |\mathcal{X}_1|, |\mathcal{X}_2|, |\mathcal{Y}|) \to 0$ such that

$$|\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y} \Rightarrow |D(\hat{P}_{X_1X_2Y}||Q_{X_1X_2Y}) - D(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y})| \leq \mu_n.$$

Hence,

$$Q^n_{X_1X_2Y}(C \times D \times F) \geq \exp[-n(D(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}) + \varsigma_n + \mu_n)], \quad (53)$$

and consequently

$$\theta(+0, +0, \epsilon) \leq \min_{\tilde{P}_{X_1X_2Y} \in \mathcal{L}_0} D(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}).$$

References

[1] W. Zhao and L. Lai, "Distributed testing against independence with conferencing encoders," in *Proc. IEEE Inform. Theory Workshop*, (Jeju Island, Korea), Oct. 2015.

[2] M.-F. Balcan, A. Blum, S. Fine, and Y. Mansour, "Distributed learning, communication complexity and privacy," in *Proc. Conf. on Learning Theory*, (Princeton, NJ), pp. 26.1–26.22, Jun. 2012.

[3] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems 26*, (Stateline, NV), pp. 2328–2336, Dec. 2013.

[4] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *Proc. Intl. Conf. on Machine Learning*, (Beijing, China), Jun. 2014.

[5] G. Katz, P. Piantanida, R. Couillet, and M. Debbah, "On the necessity of binning for the distributed hypothesis testing problem," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Hongkong, China), Jun. 2015.

[6] M. Mhanna and P. Piantanida, "On secure distributed hypothesis testing," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Hongkong, China), Jun. 2015.

[7] G. Katz, P. Piantanida, and M. Debbah, "Collaborative distributed hypothesis testing," *arXiv preprint arXiv:1604.01292*, 2016.

[8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, vol. 54, (Fort Lauderdale, FL), Apr. 2017.

[9] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inform. Theory*, vol. 32, pp. 533–542, July 1986.

[10] T. S. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2300–2324, Oct. 1998.

[11] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 33, pp. 759–772, Nov. 1987.

[12] T. Han and K. Kobayashi, "Exponential-type error probabilities for multiterminal hypothesis testing," *IEEE Trans. Inform. Theory*, vol. 35, pp. 2–14, Jan. 1989.

[13] H. Shalaby and A. Papamarcou, "Multiterminal detection with zero-rate data compression," *IEEE Trans. Inform. Theory*, vol. 38, pp. 254–267, Mar. 1992.

[14] W. Zhao and L. Lai, "Distributed detection with vector quantizer," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, pp. 105–119, Jun. 2016.

[15] M. S. Rahman and A. B. Wagner, "The optimality of binning for distributed hypothesis testing," *IEEE Trans. Inform. Theory*, vol. 58, pp. 6282–6303, Oct. 2012.

[16] W. Zhao and L. Lai, "Distributed testing with zero-rate compression," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Hongkong, China), Jun. 2015.

[17] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. Inform. Theory*, vol. 54, pp. 4666–4681, Oct. 2008.

[18] Y. Xiang and Y.-H. Kim, "Interactive hypothesis testing with communication constraints," in *Proc. Allerton Conf. on Communication, Control, and Computing*, (Montecello, IL), pp. 1065–1072, Oct. 2012.

[19] Y. Xiang and Y.-H. Kim, "Interactive hypothesis testing against independence," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Istanbul, Turkey), pp. 2840–2844, Jul. 2013.

[20] E. Akofor and B. Chen, "Interactive distributed detection: Architecture and performance analysis," *IEEE Trans. Inform. Theory*, vol. 60, pp. 6456–6473, Oct. 2014.

[21] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, UK: Cambridge Unversity Press, 2011.

[22] F. Zhu and B. Chen, "On the sum capacity of the discrete memoryless interference channel with one-sided weak interference and mixed interference," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Cambridge, MA), pp. 2271–2275, Jul. 2012.

[23] A. Kaspi, "Two-way source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 31, no. 6, pp. 735–740, 1985.

[24] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 34, no. 2, pp. 157–177, 1976.

[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1982.

Wenwen Zhao (S'15) received the B. E. degree from University of Science and Technology of China, Hefei, China in 2013.

She is currently a Ph.D. student in the Department of Electrical and Computer Engineering, University of California, Davis. Her research interests are in statistical learning and information theory.

Lifeng Lai (M'07) received the B.E. and M. E. degrees from Zhejiang University, Hangzhou, China in 2001 and 2004 respectively, and the PhD degree from The Ohio State University at Columbus, OH, in 2007. He was a postdoctoral research associate at Princeton University from 2007 to 2009, an assistant professor at University of Arkansas, Little Rock from 2009 to 2012, and an assistant professor at Worcester Polytechnic Institute. Since 2016, he has been an associate professor at University of California, Davis. Dr. Lai's research interests include information theory, stochastic signal processing and their applications in wireless communications, security and other related areas.

Dr. Lai was a Distinguished University Fellow of the Ohio State University from 2004 to 2007. He is a co-recipient of the Best Paper Award from IEEE Global Communications Conference (Globecom) in 2008, the Best Paper Award from IEEE Conference on Communications (ICC) in 2011 and the Best Paper Award from IEEE Smart Grid Communications (SmartGridComm) in 2012. He received the National Science Foundation CAREER Award in 2011, and Northrop Young Researcher Award in 2012. He served as a Guest Editor for IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing Techniques for Wireless Physical Layer Security. He is currently serving as an Editor for IEEE Transactions on Wireless Communications, and an Associate Editor for IEEE Transactions on Information Forensics and Security.