

# Distributed Hypothesis Testing with Variable-Length Coding

Sadaf Salehkalaibar, Michèle Wigger

► **To cite this version:**

Sadaf Salehkalaibar, Michèle Wigger. Distributed Hypothesis Testing with Variable-Length Coding. 18th IEEE WiOPT, Jun 2020, Volos, Greece. hal-02940595

**HAL Id: hal-02940595**

**<https://hal.archives-ouvertes.fr/hal-02940595>**

Submitted on 16 Sep 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Distributed Hypothesis Testing with Variable-Length Coding

<sup>1</sup>Sadaf Salehkalaibar and <sup>2</sup>Michèle Wigger

<sup>1</sup>ECE Department, College of Engineering, University of Tehran, Tehran, Iran, s.saleh@ut.ac.ir

<sup>2</sup>LTCI, Telecom Paris, IP Paris, 75013 Paris, France, michele.wigger@telecom-paristech.fr

**Abstract**—This paper characterizes the optimal type-II error exponent for a distributed hypothesis testing-against-independence problem when the *expected* rate of the sensor-detector link is constrained. Unlike for the well-known Ahlswede-Csiszar result that holds under a *maximum* rate constraint and where a strong converse holds, here the optimal exponent depends on the allowed type-I error exponent. Specifically, if the type-I error probability is limited by  $\epsilon$ , then the optimal type-II error exponent under an *expected* rate constraint  $R$  coincides with the optimal type-II error exponent under a *maximum* rate constraint of  $(1 - \epsilon)R$ .

## I. INTRODUCTION

Consider the distributed hypothesis testing problem in Figure 1 with a sensor and a detector observing the source sequences  $X^n$  and  $Y^n$ , and where the sensor can send a bit string  $M \in \{0, 1\}^*$  to the detector. The joint distribution depends on one of two possible hypotheses,  $\mathcal{H} = H_0$  or  $\mathcal{H} = H_1$ , and the detector has to decide based on  $Y^n$  and  $M$  which of the two hypotheses is valid. There are two error events: a type-I error indicates that the detector declares  $\hat{\mathcal{H}} = H_1$  when the correct hypothesis is  $\mathcal{H} = H_0$ , and a type-II error indicates that the detector declares  $\hat{\mathcal{H}} = H_0$  when the correct hypothesis is  $\mathcal{H} = H_1$ . The goal is to maximize the exponential decay (in the blocklength  $n$ ) of the type-II error probability under a constrained type-I error probability. The main difference of this work compared to previous works [1]–[5] is on the constraint imposed on the communication rate. While all previous works have constrained the *maximum* number of bits that the sensor can send to the detector, here we only constrain the *expected* number of bits. Our problem is thus a relaxed version of these previous works, and can be thought of as their variable-length coding counterpart.

In this paper, we specifically consider the distributed *testing-against-independence* problem introduced in [1]. In this case, under the alternative hypothesis ( $\mathcal{H} = H_1$ ) the joint distribution factorizes into the product of the marginals under the null hypothesis ( $\mathcal{H} = H_0$ ). The proposed setup can be considered as the variable-length extension of [1] thanks to the relaxed constraint on the expected number of communicated bits. The following strategy was proposed by Ahlswede and Csiszar and was shown to be optimal [1] under a maximum rate constraint. The transmitter compresses its observed source sequence  $X^n$  and describes this compressed version to the detector. If the compression fails, it sends a 0-bit to indicate this failure. The detector decides on  $\hat{\mathcal{H}} = H_1$ , whenever it receives the single 0-bit or the *joint type* (the empirical symbol

frequencies) of the compressed sequence and the observation  $Y^n$  is not close to the one expected under  $H_0$ . Otherwise it decides on  $\hat{\mathcal{H}} = H_1$ . Notice that with the described strategy, the type-I error probability can be made arbitrarily small as the blocklength  $n$  increases.

While optimal under a maximum rate constraint, a strategy with vanishing type-I error probability has to be wasteful under an expected rate constraint. The sensor should rather identify a subset of source sequences  $\mathcal{S}_n \subseteq \mathcal{X}^n$  of probability close to  $\epsilon$  and send a 0-bit whenever the observed source sequence  $X^n \in \mathcal{S}_n$ . In all other cases, the sensor should employ the Ahlswede-Csiszar strategy [1] that is optimal under the maximum rate-constraint, and so should the detector. In particular, the detector should produce  $\hat{\mathcal{H}} = H_1$  whenever it receives the single 0-bit. Compared to the Ahlswede-Csiszar strategy, this new strategy achieves the same type-II error exponent; it increases the type-I error probability by at most  $\epsilon$ ; and it has expected rate at most equal to  $(1 - \epsilon)$  times the maximum rate of the Ahlswede-Csiszar strategy.

By means of an information-theoretic converse that uses the  $\eta$ -image characterization technique of [1], [6], we show that the described strategy achieves the optimal type-II error exponent under an expected rate constraint. The optimal type-II error exponent under an expected rate constraint  $R$  coincides with the optimal exponent under a maximum rate-constraint  $(1 - \epsilon)R$ , when  $\epsilon \in (0, 1)$  denotes the allowed type-I error probability. This result implies that under an expected rate constraint the optimal type-II error exponent depends on the allowed type-I error probability and a strong converse like under a maximum rate-constraint does not hold.

## A. Notation

We mostly follow the notation in [8]. For a given pmf  $P_X$  the set of sequences whose type (symbol frequencies) is described by  $P_X$  [9] is denoted by  $\mathcal{T}^n(P_X)$ . For a given  $P_X$  and small number  $\mu > 0$ , the set of all sequences in  $\mathcal{X}^n$  whose type has  $\ell_1$ -distance from  $P_X$  at most equal to  $\mu$  is called the  $\mu$ -typical set around  $P_X$  and is denoted  $\mathcal{T}_\mu^n(P_X)$ .

For any positive integer number  $m \geq 1$ , we use  $\text{string}(m)$  to denote the bit-string of length  $\lceil \log_2(m) \rceil$  representing  $m$ . We further use sans serif font to denote bit-strings of arbitrary lengths: for example  $m$  for a deterministic bit-string and  $M$  for a random bit-string. The function  $\text{len}(m)$  returns the length of a given bit-string  $m \in \{0, 1\}^*$ . The notation  $h_b(\cdot)$  denotes the binary entropy function.

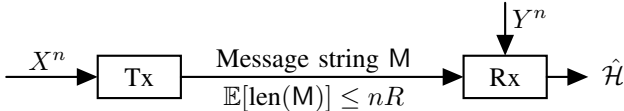


Fig. 1. Variable-length hypothesis testing.

## II. SYSTEM MODEL

Consider the distributed hypothesis testing problem with a transmitter and a receiver in Fig. 1. The transmitter observes the source sequence  $X^n$  and the receiver observes the source sequence  $Y^n$ . Under the null hypothesis

$$\mathcal{H} = H_0: (X^n, Y^n) \sim \text{i.i.d. } P_{XY}, \quad (1)$$

for a given pmf  $P_{XY}$ , whereas under the alternative hypothesis

$$\mathcal{H} = H_1: (X^n, Y^n) \sim \text{i.i.d. } P_X \cdot P_Y. \quad (2)$$

There is a noise-free bit pipe from the transmitter to the receiver. Upon observing  $X^n$ , the transmitter computes the message  $M = \phi^{(n)}(X^n)$  using a possibly stochastic encoding function

$$\phi^{(n)}: \mathcal{X}^n \rightarrow \{0, 1\}^*, \quad (3)$$

such that<sup>1</sup>

$$\mathbb{E}[\text{len}(M)] \leq nR. \quad (4)$$

It then sends a bitstring  $M$  over the bit pipe to the receiver.

The goal of the communication is that the receiver can determine the hypothesis  $\mathcal{H}$  based on its observation  $Y^n$  and its received message. Specifically, the receiver produces the guess

$$\hat{\mathcal{H}} = g^{(n)}(Y^n, M) \quad (5)$$

using a decoding function  $g^{(n)}: \mathcal{Y}^n \times \{0, 1\}^* \rightarrow \{H_0, H_1\}$ . This induces a partition of the sample space  $\mathcal{X}^n \times \mathcal{Y}^n$  into an acceptance region  $\mathcal{A}_n$  for hypothesis  $H_0$ ,

$$\mathcal{A}_n \triangleq \{(x^n, y^n): g^{(n)}(y^n, \phi^{(n)}(x^n)) = H_0\}, \quad (6)$$

and a rejection region for  $H_0$ :

$$\mathcal{A}_n^c \triangleq (\mathcal{X}^n \times \mathcal{Y}^n) \setminus \mathcal{A}_n. \quad (7)$$

*Definition 1:* For any  $\epsilon \in [0, 1)$  and for a given rate  $R \in \mathbb{R}_+$ , a type-II exponent  $\theta \in \mathbb{R}_+$  is  $(\epsilon, R)$ -achievable if there exists a sequence of functions  $(\phi^{(n)}, g^{(n)})$ , such that the corresponding sequences of type-I error probability

$$\alpha_n \triangleq P_{XY}^n(\mathcal{A}_n^c) \quad (8)$$

and type-II error probability

$$\beta_n \triangleq P_X^n P_Y^n(\mathcal{A}_n), \quad (9)$$

respectively, satisfy

$$\alpha_n \leq \epsilon, \quad (10)$$

<sup>1</sup>The expectation in (4) is with respect to the law of  $X^n$  which equals  $P_X^n$  under both hypotheses.

and

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\beta_n} \geq \theta. \quad (11)$$

The optimal exponent  $\theta_\epsilon^*(R)$  is the supremum of all  $(\epsilon, R)$ -achievable type-II exponents  $\theta \in \mathbb{R}_+$ .

## III. OPTIMAL ERROR EXPONENT

*Theorem 1:* The optimal exponent is given by

$$\theta_\epsilon^*(R) = \max_{\substack{P_{U|X}: \\ R \geq (1-\epsilon)I(U;X)}} I(U; Y). \quad (12)$$

where the mutual informations are evaluated with respect to the joint pmf

$$P_{U_{XY}} \triangleq P_{U|X} \cdot P_{XY}. \quad (13)$$

*Proof:* Here we only prove achievability. The converse is proved in Section IV.

**Achievability:** Fix a large blocklength  $n$ , a small number  $\mu \in (0, \epsilon)$ , and a conditional pmf  $P_{U|X}$  such that:

$$R = (1 - \epsilon + \mu)I(U; X) + \mu, \quad (14)$$

where the mutual information is evaluated according to the pmf in (13). Randomly generate an  $n$ -length codebook  $\mathcal{C}_U$  of rate  $R$  by picking all entries i.i.d. according to the marginal pmf  $P_U$ . The realization of the codebook

$$\mathcal{C}_U \triangleq \{u^n(m): m \in \{1, \dots, \lfloor 2^{nR} \rfloor\}\} \quad (15)$$

is revealed to all terminals.

Finally, choose a subset  $\mathcal{S}_n \subseteq \mathcal{T}_\mu^{(n)}(P_X)$  such that

$$\Pr[X^n \in \mathcal{S}_n] = \epsilon - \mu. \quad (16)$$

*Transmitter:* Assume it observes  $X^n = x^n$ . If

$$x^n \notin \mathcal{S}_n, \quad (17)$$

it looks for an index  $m$  such that

$$(u^n(m), x^n) \in \mathcal{T}_\mu^n(P_{UX}). \quad (18)$$

If successful, it picks one of these indices uniformly at random and sends the binary representation of length  $\lceil \log_2(m^*) \rceil$  of the chosen index  $m^*$  over the noiseless link:

$$M = \text{string}(m^*). \quad (19)$$

Otherwise it sends the single bit  $M = [0]$ .

*Receiver:* If it receives the single bit  $M = [0]$ , it declares  $\hat{\mathcal{H}} = H_1$ . Otherwise, it converts the received bit string  $M$  into an index  $m$  and checks whether  $(u^n(m), y^n) \in \mathcal{T}_\mu^n(P_{UY})$ . If successful, it declares  $\hat{\mathcal{H}} = H_0$ , and otherwise it declares  $\hat{\mathcal{H}} = H_1$ .

*Analysis:* Since a single bit is sent when  $x^n \in \mathcal{S}_n$  and since never more than  $n(I(U; Y) + \mu)$  bits are sent, the expected message length can be bounded as:

$$\begin{aligned} \mathbb{E}[\text{len}(M)] &= \Pr[X^n \in \mathcal{S}_n] \cdot \mathbb{E}[\text{len}(M)|X^n \in \mathcal{S}_n] \\ &\quad + \Pr[X^n \notin \mathcal{S}_n] \cdot \mathbb{E}[\text{len}(M)|X^n \notin \mathcal{S}_n] \end{aligned} \quad (20)$$

$$\leq (\epsilon - \mu) \cdot 1 + (1 - \epsilon + \mu) \cdot n(I(U; X) + \mu), \quad (21)$$

which for sufficiently large  $n$  is further bounded as (see (14)):

$$\mathbb{E}[\text{len}(\mathbf{M})] < nR. \quad (22)$$

To bound the type-I and type-II error probabilities, we notice that when  $x^n \notin \mathcal{S}_n$ , the scheme coincides with the one proposed by Ahlswede and Csiszár in [1]. When  $x^n \in \mathcal{S}_n$ , the transmitter sends the single bit  $\mathbf{M} = [0]$  and the receiver declares  $H_1$ . The type-II error probability of our scheme is thus no larger than the type-II error probability of the scheme in [1], and the type-I error probability is at most  $\Pr[X^n \in \mathcal{S}_n] = \epsilon - \mu$  larger than in [1]. Since the type-I error probability in [1] tends to 0 as  $n \rightarrow \infty$ , the type-I error probability here is bounded by  $\epsilon$ , for sufficiently large values of  $n$  and all choices of  $\mu \in (0, \epsilon)$ . Combining the result in [1], with (22), and letting  $\mu \rightarrow 0$  thus establishes the achievability part of the proof. For the converse proof see the following Section IV. ■

#### IV. PROOF OF CONVERSE TO THEOREM 1

Fix an achievable exponent  $\theta < \theta_\epsilon^*(R)$  and a sequence of encoding and decision functions so that (10) and (11) are satisfied. Fix also an integer  $n$  and a small number  $\eta \geq 0$  and define the set

$$\mathcal{B}_n(\eta) \triangleq \left\{ x^n : \Pr \left[ \hat{\mathcal{H}} = H_0 \mid X^n = x^n, \mathcal{H} = H_0 \right] \geq \eta \right\}. \quad (23)$$

Notice that by the constraint on the type-I error probability, (10),

$$\begin{aligned} 1 - \epsilon &\leq \sum_{x^n \in \mathcal{B}_n(\eta)} \Pr \left[ \hat{\mathcal{H}} = H_0 \mid X^n = x^n, \mathcal{H} = H_0 \right] P_X^n(x^n) \\ &\quad + \sum_{x^n \notin \mathcal{B}_n(\eta)} \Pr \left[ \hat{\mathcal{H}} = H_0 \mid X^n = x^n, \mathcal{H} = H_0 \right] P_X^n(x^n) \end{aligned} \quad (24)$$

$$\leq P_X^n(\mathcal{B}_n(\eta)) + \eta(1 - P_X^n(\mathcal{B}_n(\eta))). \quad (25)$$

Thus,

$$P_X^n(\mathcal{B}_n(\eta)) \geq \frac{1 - \epsilon - \eta}{1 - \eta}. \quad (26)$$

Define now

$$\mu_n \triangleq n^{-\frac{1}{3}} \quad (27)$$

and

$$\mathcal{D}_n(\eta) \triangleq \mathcal{T}_{\mu_n}^n(P_X) \cap \mathcal{B}_n(\eta). \quad (28)$$

By [10, Lemma 2.12]:

$$P_X^n(\mathcal{T}_{\mu_n}^n(P_X)) \geq 1 - \frac{|\mathcal{X}|}{2\mu_n n}, \quad (29)$$

which combined with (26) and the general identity  $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$  yields:

$$P_X^n(\mathcal{D}_n(\eta)) \geq \frac{1 - \epsilon - \eta}{1 - \eta} - \frac{|\mathcal{X}|}{2\mu_n n} \triangleq \Delta_n. \quad (30)$$

Define the random variables  $(\tilde{\mathbf{M}}, \tilde{X}^n, \tilde{Y}^n)$  as the restriction of the triple  $(\mathbf{M}, X^n, Y^n)$  to  $X^n \in \mathcal{D}_n(\eta)$ . The probability distribution of the restricted triple is then given by:

$$\begin{aligned} P_{\tilde{\mathbf{M}}\tilde{X}^n\tilde{Y}^n}(m, x^n, y^n) &\triangleq \\ P_{XY}^n(x^n, y^n) \cdot \frac{\mathbb{1}\{x^n \in \mathcal{D}_n(\eta)\}}{P_X^n(\mathcal{D}_n(\eta))} \cdot \mathbb{1}\{\phi^{(n)}(x^n) = m\}. \end{aligned} \quad (31)$$

This implies in particular:

$$P_{\tilde{X}^n}(x^n) \leq P_X^n(x^n) \cdot \Delta_n^{-1}, \quad (32)$$

$$P_{\tilde{Y}^n}(y^n) \leq P_Y^n(y^n) \cdot \Delta_n^{-1}, \quad (33)$$

$$P_{\tilde{\mathbf{M}}}(m) \leq P_{\mathbf{M}}(m) \cdot \Delta_n^{-1}, \quad (34)$$

and

$$D(P_{\tilde{X}^n} \| P_X^n) \leq \log \Delta_n^{-1}. \quad (35)$$

Single-letter characterization of the rate constraint: Define the random variables  $L \triangleq \text{len}(\mathbf{M})$  and  $\tilde{L} \triangleq \text{len}(\tilde{\mathbf{M}})$ , and notice that by the rate constraint (4):

$$nR \geq \mathbb{E}[L] \quad (36)$$

$$\begin{aligned} &= \mathbb{E}[L | X^n \in \mathcal{D}_n(\eta)] \cdot P_X^n(\mathcal{D}_n(\eta)) \\ &\quad + \mathbb{E}[L | X^n \notin \mathcal{D}_n(\eta)] \cdot (1 - P_X^n(\mathcal{D}_n(\eta))) \end{aligned} \quad (37)$$

$$\geq \mathbb{E}[L | X^n \in \mathcal{D}_n(\eta)] \cdot P_X^n(\mathcal{D}_n(\eta)) \quad (38)$$

$$= \mathbb{E}[\tilde{L}] \cdot P_X^n(\mathcal{D}_n(\eta)) \quad (39)$$

$$\geq \mathbb{E}[\tilde{L}] \cdot \Delta_n, \quad (40)$$

where (39) holds because  $\tilde{\mathbf{M}}$  is obtained by restricting  $\mathbf{M}$  to the event  $X^n \in \mathcal{D}_n(\eta)$  and  $\tilde{L}$  denotes the length of  $\tilde{\mathbf{M}}$ ; and step (40) holds by the definition of  $\Delta_n$  in (30).

Now, since  $\tilde{L}$  is function of  $\tilde{\mathbf{M}}$ , we have:

$$H(\tilde{\mathbf{M}}) = H(\tilde{\mathbf{M}}, \tilde{L}) \quad (41)$$

$$= H(\tilde{\mathbf{M}} | \tilde{L}) + H(\tilde{L}) \quad (42)$$

$$= \sum_{\ell} \Pr(\tilde{L} = \ell) H(\tilde{\mathbf{M}} | \tilde{L} = \ell) + H(\tilde{L}) \quad (43)$$

$$\leq \sum_{\ell} \Pr(\tilde{L} = \ell) \ell + H(\tilde{L}) \quad (44)$$

$$= \mathbb{E}[\tilde{L}] + H(\tilde{L}) \quad (45)$$

$$\leq \frac{nR}{\Delta_n} + H(\tilde{L}) \quad (46)$$

$$\leq \frac{nR}{\Delta_n} + \frac{nR}{\Delta_n} h_b \left( \frac{\Delta_n}{nR} \right) \quad (47)$$

$$= \frac{nR}{\Delta_n} \left( 1 + h_b \left( \frac{\Delta_n}{nR} \right) \right). \quad (48)$$

Here, (44) holds because when  $\mathbf{M}$  consists of  $\ell$  bits ( $L = \ell$ ), then its entropy cannot exceed  $\ell$ ; (46) follows from (40); and (47) holds because when  $\mathbb{E}[\tilde{L}] \leq \frac{nR}{\Delta_n}$ , then the entropy of  $\tilde{L}$

can be at most that of a Geometric distribution with mean  $\frac{nR}{\Delta_n}$ , which is  $\frac{nR}{\Delta_n} \cdot h_b\left(\frac{\Delta_n}{nR}\right)$ .

On the other hand, we can lower bound  $H(\tilde{M})$  in the following way:

$$H(\tilde{M}) \geq I(\tilde{M}; \tilde{X}^n) \quad (49)$$

$$= H(\tilde{X}^n) - H(\tilde{X}^n | \tilde{M}) \quad (50)$$

$$= - \sum_{x^n} P_{\tilde{X}^n}(x^n) \log P_{\tilde{X}^n}(x^n) - H(\tilde{X}^n | \tilde{M}) \quad (51)$$

$$\geq - \sum_{x^n} P_{\tilde{X}^n}(x^n) \log P_{X^n}(x^n) + \log \Delta_n - H(\tilde{X}^n | \tilde{M}) \quad (52)$$

$$= - \sum_{x^n} P_{\tilde{X}^n}(x^n) \sum_{t=1}^n \log P_X(x_t) + \log \Delta_n - H(\tilde{X}^n | \tilde{M}) \quad (53)$$

$$= - \sum_{t=1}^n \sum_{x_t} P_{\tilde{X}_t}(x_t) \log P_X(x_t) + \log \Delta_n - H(\tilde{X}^n | \tilde{M}) \quad (54)$$

$$= \sum_{t=1}^n H(\tilde{X}_t) + \sum_{t=1}^n D(P_{\tilde{X}_t} \| P_X) + \log \Delta_n - H(\tilde{X}^n | \tilde{M}) \quad (55)$$

$$= \sum_{t=1}^n \left[ H(\tilde{X}_t) - H(\tilde{X}_t | \tilde{M}, \tilde{X}^{t-1}) \right] + \sum_{t=1}^n D(P_{\tilde{X}_t} \| P_X) + \log \Delta_n \quad (56)$$

$$= \sum_{t=1}^n I(\tilde{U}_t; \tilde{X}_t) + \sum_{t=1}^n D(P_{\tilde{X}_t} \| P_X) + \log \Delta_n \quad (57)$$

$$= nI(\tilde{U}_T; \tilde{X}_T | T) + \sum_{t=1}^n \sum_{x \in \mathcal{X}} P_{\tilde{X}_T | T=t}(x) \log \frac{P_{\tilde{X}_T | T=t}(x)}{P_X(x)} + \log \Delta_n \quad (58)$$

$$= nI(\tilde{U}_T; \tilde{X}_T | T) + \sum_{t=1}^n \sum_{x \in \mathcal{X}} P_{\tilde{X}_T | T=t}(x) \log \frac{P_{\tilde{X}_T | T=t}(x)}{P_{\tilde{X}_T}(x)} + \sum_{t=1}^n \sum_{x \in \mathcal{X}} P_{\tilde{X}_T | T=t}(x) \log \frac{P_{\tilde{X}_T}(x)}{P_{X_t}(x)} + \log \Delta_n \quad (59)$$

$$= nI(\tilde{U}_T; \tilde{X}_T | T) + nI(\tilde{X}_T; T) + nD(P_{\tilde{X}_T} \| P_{X_T}) + \log \Delta_n \quad (60)$$

$$\geq nI(\tilde{U}_T, T; \tilde{X}_T) + \log \Delta_n \quad (61)$$

$$= nI(U; \tilde{X}) + \log \Delta_n, \quad (62)$$

where

- (52) holds by (32);
- (53) holds because  $X^n$  is i.i.d. under  $P_X^n$ ;
- (57) holds by defining  $\tilde{U}_t \triangleq (\tilde{M}, \tilde{X}^{t-1})$ ;

- (60) holds because  $T$  is uniformly chosen over  $\{1, \dots, n\}$ ;
- (62) follows by defining  $U \triangleq (\tilde{U}_T, T)$  and  $\tilde{X} \triangleq \tilde{X}_T$ .

Combining (48) and (62), we obtain:

$$R \geq \frac{I(U; \tilde{X}) + \frac{1}{n} \log \Delta_n}{1 + h_b\left(\frac{\Delta_n}{nR}\right)} \cdot \Delta_n. \quad (63)$$

*Upper bounding the error exponent:* For each string  $m \in \{0, 1\}^*$ , define the following sets:

$$\mathcal{F}_m \triangleq \left\{ x^n \in \mathcal{X}^n : \phi^{(n)}(x^n) = m \right\} \cap \mathcal{D}_n(\eta), \quad (64)$$

$$\mathcal{G}_m \triangleq \left\{ y^n \in \mathcal{Y}^n : g^{(n)}(y^n, m) = H_0 \right\}. \quad (65)$$

Using (34), the type-II error probability can then be lower bounded as:

$$\beta_n = \sum_m P_M(m) \cdot P_Y^n(\mathcal{G}_m) \geq \Delta_n \cdot \sum_m P_{\tilde{M}}(m) \cdot P_Y^n(\mathcal{G}_m). \quad (66)$$

In order to find a lower bound to the right hand-side of (66), we need the following definition and lemma. A set  $\mathcal{B} \subseteq \mathcal{Y}^n$  is an  $\eta$ -image of the set  $\mathcal{A} \subseteq \mathcal{X}^n$  if

$$P_{Y|X}^n(\mathcal{B} | x^n) \geq \eta, \quad \forall x^n \in \mathcal{A}. \quad (67)$$

The following lemma is a simple restatement of the lemma proved in [6].

*Lemma 1 (Lemma 3 in [6]):* Consider a set  $\mathcal{A} \subseteq \mathcal{X}^n$ , a number  $\eta \in (0, 1)$ , and an  $\eta$ -image  $\mathcal{B}$  of  $\mathcal{A}$  with respect to the channel  $P_{Y|X}$ . Then, for any number  $\delta' > 0$  and any output distribution  $P_{Y_A}^n$  induced over the channel  $P_{Y|X}^n$  by an arbitrary input distribution  $P_A$  on  $\mathcal{A}$ , i.e.,

$$P_{Y_A}^n(y^n) \triangleq \sum_{x^n \in \mathcal{A}} P_A(x^n) P_{Y|X}^n(y^n | x^n), \quad (68)$$

for all sufficiently large blocklengths  $n$ :

$$P_{Y^n}(B) \geq 2^{-D(P_{Y_A} \| P_Y^n) - n\delta'}. \quad (69)$$

To apply this lemma, we notice that the set  $\mathcal{G}_m$  is an  $\eta$ -image of the set  $\mathcal{F}_m$ . In fact, by (23), under  $\mathcal{H} = H_0$ , whenever  $X^n \in \mathcal{D}_n(\eta)$  the receiver guesses  $\hat{\mathcal{H}} = H_0$  with probability at least  $\eta$ . Since  $X^n \in \mathcal{F}_m$  implies  $X^n \in \mathcal{D}_n(\eta)$  and  $M = m$ , the probability that  $Y^n \in \mathcal{G}_m$  needs to be at least  $\eta$ .

We can use this observation and Lemma 1 to further lower bound the sum in (66) for any  $\delta' > 0$  and any sufficiently large  $n$ :

$$\begin{aligned} & \sum_m P_{\tilde{M}}(m) \cdot P_Y^n(\mathcal{G}_m) \\ & \geq 2^{-n\delta'} \sum_m P_{\tilde{M}}(m) 2^{-D(P_{\tilde{Y}^n | \tilde{M}=m} \| P_Y^n)} \end{aligned} \quad (70)$$

$$\geq 2^{-n\delta'} 2^{-\sum_m P_{\tilde{M}}(m) D(P_{\tilde{Y}^n | \tilde{M}=m} \| P_Y^n)} \quad (71)$$

where

- (70) holds by Lemma 1 for the choice  $A = \mathcal{F}_m$ , because  $\mathcal{G}_m$  is an  $\eta$ -image of the set  $\mathcal{F}_m$  and because according to (31),  $P_{\tilde{Y}^n|\tilde{M}}(\cdot|m)$  is the output distribution induced by channel  $P_{\tilde{Y}|X}^n$  for input distribution  $P_{\tilde{X}^n|\tilde{M}}(\cdot|m)$  over the set  $\mathcal{F}_m$ ;
- (71) holds by the convexity of the function  $t \mapsto 2^t$ .

We define  $\delta'' \triangleq \delta' - \frac{1}{n} \log \Delta_n$  and combine (66) with (71) to obtain:

$$-\frac{1}{n} \log \beta_n \leq \frac{1}{n} \sum_m \sum_{y^n \in \mathcal{Y}^n} P_{\tilde{M}\tilde{Y}^n}(m, y^n) \log \frac{P_{\tilde{Y}^n|\tilde{M}}(y^n|m)}{P_{\tilde{Y}}^n(y^n)} + \delta'' \quad (72)$$

$$= \frac{1}{n} D(P_{\tilde{M}\tilde{Y}^n} \| P_{\tilde{M}} P_{\tilde{Y}}^n) + \delta'' \quad (73)$$

$$= \frac{1}{n} D(P_{\tilde{M}\tilde{Y}^n} \| P_{\tilde{M}} P_{\tilde{Y}^n}) + \frac{1}{n} E_{P_{\tilde{Y}^n}} \left[ \log \frac{P_{\tilde{Y}^n}}{P_{\tilde{Y}}^n} \right] + \delta'' \quad (74)$$

$$\leq \frac{1}{n} D(P_{\tilde{M}\tilde{Y}^n} \| P_{\tilde{M}} P_{\tilde{Y}^n}) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (75)$$

$$= \frac{1}{n} I(\tilde{M}; \tilde{Y}^n) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (76)$$

$$= \frac{1}{n} \sum_{t=1}^n I(\tilde{M}; \tilde{Y}_t | \tilde{Y}^{t-1}) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (77)$$

$$\leq \frac{1}{n} \sum_{t=1}^n I(\tilde{M}, \tilde{Y}^{t-1}; \tilde{Y}_t) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (78)$$

$$\leq \frac{1}{n} \sum_{t=1}^n I(\tilde{M}, \tilde{X}^{t-1}; \tilde{Y}_t) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (79)$$

$$= \frac{1}{n} \sum_{t=1}^n I(\tilde{U}_t; \tilde{Y}_t) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (80)$$

$$= I(\tilde{U}_T; \tilde{Y}_T | T) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (81)$$

$$\leq I(\tilde{U}_T, T; \tilde{Y}_T) + \frac{1}{n} \log \Delta_n^{-1} + \delta'' \quad (82)$$

$$= I(U; \tilde{Y}) + \frac{1}{n} \log \Delta_n^{-1} + \delta'', \quad (83)$$

where

- (75) holds by (33) and (34);
- (79) holds by the Markov chain  $\tilde{Y}^{t-1} \rightarrow (\tilde{M}, \tilde{X}^{t-1}) \rightarrow \tilde{Y}_t$ ;
- (83) follows by defining  $\tilde{Y} \triangleq \tilde{Y}_T$ .

Thus, from (83), we have:

$$-\frac{1}{n} \log \beta_n \leq I(U; \tilde{Y}) + \frac{1}{n} \log \Delta_n^{-1} + \delta''. \quad (84)$$

Notice that according to (31), the distribution  $P_{\tilde{X}^n}$  is a restriction to the set  $\mathcal{D}_n(\eta)$  which is a subset of the typical set, thus we have  $|P_{\tilde{X}} - P_X| \leq \mu_n$ . Also, from  $P_{\tilde{Y}|\tilde{X}} = P_{Y|X}$ , and by the uniform continuity of the involved information quantities, we get as  $n \rightarrow \infty$  and  $\eta \rightarrow 0$ :

$$R \geq (1 - \epsilon)I(U; X), \quad (85)$$

$$\theta \leq I(U; Y). \quad (86)$$

This concludes the proof of the converse.

## V. CONCLUSION

We established the optimal type-II error exponent of a distributed testing-against-independence problem under a constraint on the probability of type-I error and on the expected communication rate. This result can be seen as a variable-length coding version of the well-known result by Ahlswede and Csiszar [1] which holds under a maximum rate-constraint. Interestingly, the optimal type-II error exponent under an expected rate constraint  $R$  coincides with the optimal type-II error exponent under a maximum rate constraint  $(1 - \epsilon)R$  when the type-I error probability is constrained to be at most  $\epsilon \in (0, 1)$ . Thus, unlike in the scenario with a maximum rate constraint, here the strong converse fails because the optimal type-II error exponent depends on the allowed type-I error probability  $\epsilon$ .

## ACKNOWLEDGEMENTS

M. Wigger and S. Salehkalaibar acknowledge funding support from the ERC under grant agreement 715111.

## REFERENCES

- [1] R. Ahlswede and I. Csiszar, "Hypothesis testing with communication constraints," *IEEE Trans. on Info. Theory*, vol. 32, pp. 533–542, Jul. 1986.
- [2] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. on Info. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987.
- [3] H. Shimokawa, T. Han, and S. I. Amari, "Error bound for hypothesis testing with data compression," in *Proc. IEEE Int. Symp. on Info. Theory*, Jul. 1994, p. 114.
- [4] M. S. Rahman and A. B. Wagner, "On the optimality of binning for distributed hypothesis testing," *IEEE Trans. on Info. Theory*, vol. 58, no. 10, pp. 6282–6303, Oct. 2012.
- [5] N. Weinberger and Y. Kochman, "On the reliability function of distributed hypothesis testing under optimal detection," *IEEE Trans. on Info. Theory*, 2019.
- [6] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. on Info. Theory*, vol. 54, no. 10, pp. 4666–4681, Oct. 2008.
- [7] H. Tyagi and S. Watanabe, "Strong converse using change of measure arguments," 2018. [Online]. Available: <https://arxiv.org/pdf/1805.04625.pdf>
- [8] A. El Gamal and Y. H. Kim, *Network Information Theory*. Cambridge University Press, 2011.
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Ed.* Wiley, 2006.
- [10] I. Csiszar and J. Korner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.