# Successive Refinement for Hypothesis Testing and Lossless One-helper Problem

Chao Tian, *Member, IEEE*, Jun Chen, *Member, IEEE*

***Abstract*—We investigate two closely related successive refinement (SR) coding problems: (i) In the hypothesis testing (HT) problem, bivariate hypothesis $H_0 : P_{XY}$ against $H_1 : P_X P_Y$, i.e., test against independence is considered. One remote sensor collects data stream $X$ and sends summary information, constrained by SR coding rates, to a decision center which observes data stream $Y$ directly. (ii) In the one-helper (OH) problem, $X$ and $Y$ are encoded separately and the receiver seeks to reconstruct $Y$ losslessly. Multiple levels of coding rates are allowed at the two sensors, and the transmissions are performed in an SR manner.**

**We show that the SR-HT rate-error-exponent region and the SR-OH rate region can be reduced to essentially the same entropy characterization form. Single-letter solutions are thus provided in a unified fashion, and the connection between them is discussed. These problems are also related to the information bottleneck (IB) problem, and through this connection we provide a straightforward operational meaning for the IB method. Connection to the pattern recognition problem, the notion of successive refinability, and two specific sources are also discussed. A strong converse for the SR-HT problem is proved by generalizing the image size characterization method, which shows the optimal type-two error exponents under constant type-one error constraints are independent of the exact values of those constants.**

***Index Terms*—Entropy characterization, error exponent, hypothesis testing, image size characterization, information bottleneck, one-helper problem, successive refinement.**

## I. Introduction

In conventional successive refinement (SR) source coding, a source stream is encoded into more than one description in a progressive order such that later descriptions can be used to refine the early ones, resulting in progressive reconstructions of improving qualities. As such, it can be conveniently formulated as a rate-distortion problem. In addition to the fundamental problem of characterizing the rate-distortion region, also of interest is the condition under which such a progressive coding requirement does not cause any performance loss, compared to a single stage coding system. These questions were the focus of early works [1]–[3]. The rate-distortion problem with various extensions has subsequently been thoroughly researched, among which are the notable work by Effros [4], [5] and by Tuncel and Rose [6]–[8].

The successive refinement coding structure is clearly appealing in multimedia delivery systems, since such a framework allows a single copy of the multimedia content on the server to satisfy requirement by users with different communication capabilities. However, the importance of successive refinement coding goes well beyond this single specific application, and in the present work we investigate several such cases which deviate from the traditional rate-distortion setting. In the remainder of this section, we review related previous work on the hypothesis testing (HT) problem and the one-helper (OH) problem; the successive refinement version of these problems in consideration and our contribution are also outlined. Formal problem definitions are given in the next section.

Chao Tian was with School of Computer and Communication Science, Ecole Polytechnique Federale de Lausanne, Lausanne, CH1015, Switzerland. He is now with AT&T Labs–Research, Florham Park, NJ 07932.

Jun Chen is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada L8S 4K1 (email: junchen@ece.mcmaster.ca).

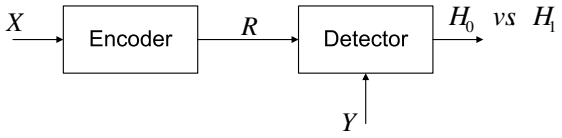


Fig. 1. Hypothesis testing with one remote sensor.

### A. The hypothesis testing problem

The information theoretic formulation of the hypothesis testing problem under communication constraint first appeared in the award-winning article by Ahlswede and Csiszár [9], and the problem can be described as follows (see also Fig. 1). Source stream $X$ is observed by a remote sensor who communicates to the receiver under certain rate constraint $R < H(X)$, and the receiver, which observes another dependent source stream $Y$, wishes to distinguish between the two hypotheses $H_0 : P_{XY}$ and $H_1 : Q_{XY}$. The problem is to characterize the exponent of the type-two error ($H_1$ is true but the detector judges otherwise), when the type-one error ($H_0$ is true but the detector judges otherwise) is less than a pre-specified probability $\epsilon$.

For the case that $Q_{XY} = P_X P_Y$, i.e., testing against independence, single letter characterization of the error exponent was given in [9] for an arbitrary $\epsilon \in (0, 1)$. This is the equivalence of the "strong converse" result encountered in Shannon theory as pointed out by Ahlswede and Csiszár, in comparison to the "weak converse" for which only the case $\epsilon \to 0$ is considered. For a general alternative hypothesis $Q_{XY}$, single letter lower and upper bounds were provided, yet a complete characterization was not found. Many subsequent works extended or strengthened the results in [9], for example, when both sensors are remote, or when type-one error is constrained to satisfy certain error exponent requirement. The review article by Han and Amari [10] provides a comprehensive summary of literature on this topic.

In this work we consider the same distributed setting as in [9] with one remote sensor, however, the receiver, instead of waiting for the completion of the rate $R$ transmission to make a single decision in the end, wishes to form a preliminary

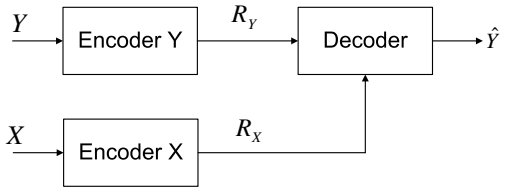


Fig. 2. Lossless one-helper problem.

decision based on a subset of the description, and may (or may not) wait for the completion of the transmission to form a final decision. The remote sensor encoder thus has to take this requirement into consideration. This problem clearly has practical meaning in distributed hypothesis testing system with delay constraint, and will be referred to as the successive refinement hypothesis testing (SR-HT) problem.

We shall focus on the testing against independence case in this work, and provide a single letter characterization of the rate-error-exponent region. Furthermore, it is shown that the result holds independent of the exact value of the constant type-one error constraints, i.e., the strong converse result is established. Interestingly, when the type-one error constraints are sufficiently large, it can be shown that the progressive encoding requirement does not cause any performance loss compared to single stage coding, in terms of type-two error exponent. It is worth mentioning that the proof for the strong converse is not a trivial generalization of the proof in [9]. It appears that the covering lemma in [11], which is an important tool in proving the strong converse for the single stage case, is not sufficient for the successive refinement setting. To circumvent this difficulty, we generalize the image size characterization method [12] to provide the desired proof.

### *B. The one-helper source coding problem*

The lossless one-helper (OH) source coding problem was considered independently by Wyner [13] and by Ahlswede and Körner [14], which can be described as follows (see also Fig. 2). Two correlated source $X$ and $Y$ observed by two sensors are encoded separately into descriptions of rate $R_X$ and $R_Y$, respectively. The decoder wishes to reconstruct $Y$ losslessly based on information received from both sensors. A conclusive result was provided in [13][14] for the achievable rate region of this problem. The lossy version of the one-helper problem is more difficult, for which the only solved special case is the Gaussian source problem under the quadratic distortion measure [15].

We extend the above lossless one-helper problem to the successive refinement setting (referred to as the SR-OH problem). Note that in this extension the requirement on the reconstruction is still lossless, but the encoding is done in an SR fashion, and thus the decoder receives SR information regarding the source from either of the two encoders; we believe this is a natural generalization of the SR notion from the conventional rate-distortion setting. Though in this work we mainly use this problem as an "enabler" to the hypothesis testing problem, it is indeed well motivated in practice. Observe that in the original problem, the two sources are encoded and transmitted separately to the receiver. As such one particular sensor encoder might not have accurate information as to what the capacity of the communication link is between the receiver and the other sensor, or even whether the other link is reliable or not. If the link between one sensor and the receiver fails after certain amount of data is successfully transmitted, the data from the other sensor will not be sufficient for the receiver to recover from this failure, when the existing coding scheme for the OH problem [13][14] is used. One solution is that instead of fixing one final operating point $(R_X, R_Y)$, the sensors choose several possible operating rate pairs and the information is transmitted progressively, such that as long as the received information from both sensors is sufficient jointly, the decoding procedure can be performed. In the situation described above, the refinement information from the other sensor with working communication link can then compensate for the lost information. This approach is also applicable when one of the communication links suffers unexpected delay or degradation of quality, and the other sensor with working link can help reduce this delay by sending additional information. In a sense, this successive refinement coding structure makes the system more robust to communication link failure; problems in a similar vein can be found in [16] and [17]. In this work, we shall show that the achievable rate region for the SR-OH problem has essentially the same entropy characterization form as that of the SR-HT problem, and also provide a conclusive single-letter solution for this problem.

### *C. Motivation and structure of the paper*

In addition to the clear application of the two problems which have not been treated before in the literature, one of our main motivations is that these problems are closely related and it is beneficial to make a unified investigation of them. The connection has been recognized for the single stage case in [9], and we show that it continues to hold for the successive refinement case. In fact, it appears difficult to establish the direct half of the hypothesis testing problem directly, but through this relation the proof is rather straightforward, which is exactly the approach taken in [9]. It will also be shown that a single codebook exists which is good for these problems. Furthermore, existing results in one problem can be readily applied to the other problem to give rather non-trivial results. For example, the successive refinability of the doubly symmetric binary source for the hypothesis testing can be derived directly from a result by Wyner [18].

These two problem are related to the pattern recognition problem [19]–[23] and the information bottleneck problem [24]. In fact, the entropy characterization problem extracted from the problems being considered also readily provides an operational meaning for the information bottleneck (IB) method [24]. Though several attempts were made to formalize and clarify the operational meaning of the IB function [25], [26], our approach is more straightforward and intuitive. This shows the importance of the IB method, as it is not merely useful as a classification tool [27], but has roots in many information theoretic problems.

The Gaussian source is given special consideration, and it is shown that lattice encoding together with an approximation to the Neyman-Pearson detector, namely the weighed distance difference detector, is asymptotically optimal for this problem. Large deviation technique is used to establish this result.

The rest of the paper is organized as follows. In Section II we provide formal definitions for the problems. In Section III the main results are presented. In IV the concept of successive refinability is defined, and sufficient and necessary conditions are provided. The doubly symmetric binary source is investigated in this context. In Section V the Gaussian source is considered and we provide a lattice approach for this case. Section VI gives the strong converse proof for the hypothesis testing problem. Finally Section VII concludes the paper.

## II. Notation and Preliminaries

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets. Let $\mathcal{X}^n$ be the set of all $n$-vectors with components in $\mathcal{X}$. Denote an arbitrary member of $\mathcal{X}^n$ as $x^n = (x_1, x_2, ..., x_n)$, or alternatively as $\boldsymbol{x}$. Upper case is used for random variables and vectors. A discrete memoryless source (DMS) $(\mathcal{X}, P_X)$ is an infinite sequence $\{X_i\}_{i=1}^{\infty}$ of independent copies of a random variable $X$ in $\mathcal{X}$ with a generic distribution $P_X$ and $P_X(x^n) = \prod_{i=1}^n P_X(x_i)$. Similarly, let $(\mathcal{X}, \mathcal{Y}, P_{XY})$ be a discrete memoryless two-source with generic distribution $P_{XY}$; the subscript will be dropped when it is clear from the context as $P(X, Y)$. Without loss of generality, we assume $P_X(x) \neq 0$ for any $x \in \mathcal{X}$ and similarly for $P_Y$. The cardinality of a set $S$ is denoted as $|S|$.

In this work only two stage systems will be considered. To distinguish between the two problems when necessary, the subscripts "ht" and "oh" are used for "**H**ypothesis **T**esting" and "**O**ne-**H**elper", respectively.

### A. Successive refinement for hypothesis testing

Let the two hypotheses be given as follows

$$\begin{aligned} H_0 &: P_{XY} = (P_{XY}(x,y))_{x\in\mathcal{X},y\in\mathcal{Y}}, \\ H_1 &: Q_{XY} = P_X \times P_Y = (P_X(x)P_Y(y))_{x\in\mathcal{X},y\in\mathcal{Y}}, \end{aligned}$$

where $P_X$ and $P_Y$ are the marginal distributions of $P_{XY}$. In other words, we are to test against independence.

***Definition** 1:* An $(n, \epsilon_1, \epsilon_2, \beta_1, \beta_2, M_1, M_2)$ SR-HT code consists of two encoding functions

$$f_1 : X^n \to I_{M_1}, \quad f_2 : X^n \to I_{M_2}, \tag{1}$$

where $I_M = \{1, 2, ..., M\}$ and two detectors specified by the decision set $A_1 \subseteq I_{M_1} \times \mathcal{Y}^n$ and $A_2 \subseteq I_{M_1} \times I_{M_2} \times \mathcal{Y}^n$ as:

$$g_{t,1}(i_1, y^n) = \begin{cases} H_0 & (i_1, y^n) \in A_1; \\ H_1 & \text{otherwise}, \end{cases}$$

$$g_{t,2}(i_1, i_2, y^n) = \begin{cases} H_0 & (i_1, i_2, y^n) \in A_2; \\ H_1 & \text{otherwise}, \end{cases}$$

such that the type-one errors at the two stages do not exceed fixed $\epsilon_1, \epsilon_2 \in (0, 1)$, respectively; i.e.,

$$\begin{aligned} P_{f_1(X^n)Y^n}(A_1) &\geq 1 - \epsilon_1, \\ P_{f_1(X^n)f_2(X^n)Y^n}(A_2) &\geq 1 - \epsilon_2, \end{aligned}$$

and the type-two errors at the two stages do not exceed $\beta_1, \beta_2$, respectively; i.e.,

$$\begin{aligned} Q_{XY}(A_1) &= P_{f_1(X^n)} \times P_{Y^n}(A_1) \leq \beta_1, \\ Q_{XY}(A_2) &= P_{f_1(X^n)f_2(X^n)} \times P_{Y^n}(A_2) \leq \beta_2. \end{aligned}$$

***Definition** 2 (Achievable rates-error exponents):* A rate and type-two error exponent quadruple $(R_1, R_2, E_1, E_2)$ is said to be $(\epsilon_1, \epsilon_2)$-achievable with fixed $\epsilon_1, \epsilon_2 \in (0, 1)$, if for any $\epsilon > 0$ and sufficiently large $n$, there exists an $(n, \epsilon_1, \epsilon_2, \beta_1, \beta_2, M_1, M_2)$ SR-HT code such that

$$\begin{aligned} \frac{1}{n}\log M_1 &\leq R_1 + \epsilon, & \frac{1}{n}\log M_2 &\leq R_2 + \epsilon, \\ -\frac{1}{n}\log \beta_1 &\geq E_1 - \epsilon, & -\frac{1}{n}\log \beta_2 &\geq E_2 - \epsilon. \end{aligned}$$

Denote all the $(\epsilon_1, \epsilon_2)$-achievable quadruple as $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$, and this is the region we seek to characterize. Clearly we have $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2) \subseteq \mathcal{R}_{ht}(\epsilon_1', \epsilon_2')$ if $\epsilon_1 \leq \epsilon_1'$, $\epsilon_1 \leq \epsilon_2'$, and thus the following limit is well-defined.

***Definition** 3:* The *weakly achievable* rate-error-exponent region $\mathcal{R}_{ht}$ is

$$\mathcal{R}_{ht} \triangleq \bigcap_{\epsilon_1>0, \epsilon_2>0} \mathcal{R}_{ht}(\epsilon_1, \epsilon_2).$$

In Section VI we show that the strong converse holds true that $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$ is essentially independent of $(\epsilon_1, \epsilon_2)$, and thus a characterization of $\mathcal{R}_{ht}$ is almost a sufficient characterization of $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$.

For convenience, define the error-exponent-rate function $E(R)$ as the single-stage achievable error exponent with rate no larger than $R$, which was shown in [9] to be

$$\begin{aligned} E(R) = \max_U \{ & I(U;Y) | U \leftrightarrow X \leftrightarrow Y, \\ & I(X;U) \leq R, |\mathcal{U}| \leq |\mathcal{X}| + 1\}. \end{aligned} \tag{2}$$

As shown in [9], $E(R)$ is independent of the type-one error constraint taken value in $(0, 1)$.

### B. Successive refinement for the one-helper problem

***Definition** 4:* An $(n, M_1, M_2, M_{Y,1}, M_{Y,2}, \Delta_1, \Delta_2)$ SR-OH code for source $(\mathcal{X}, \mathcal{Y}, P_{XY})$ consists of four encoding functions

$$\begin{aligned} f_1 &: \mathcal{X}^n \to I_{M_1}, & f_2 &: \mathcal{X}^n \to I_{M_2}, \\ f_{Y,1} &: \mathcal{Y}^n \to I_{M_{Y,1}}, & f_{Y,2} &: \mathcal{Y}^n \to I_{M_{Y,2}}, \end{aligned}$$

and two decoding functions

$$\begin{aligned} g_{h,1} &: I_{M_1} \times I_{M_{Y,1}} \times I_{M_{Y,2}} \to \mathcal{Y}^n, \\ g_{h,2} &: I_{M_1} \times I_{M_2} \times I_{M_{Y,1}} \to \mathcal{Y}^n, \end{aligned}$$

such that

$$\begin{aligned} \Pr(Y^n \neq g_{h,1}(f_1(X^n), f_{Y,1}(Y^n), f_{Y,2}(Y^n))) &\leq \Delta_1, \\ \Pr(Y^n \neq g_{h,2}(f_1(X^n), f_2(X^n), f_{Y,1}(Y^n))) &\leq \Delta_2. \end{aligned}$$

***Definition** 5:* A rate quadruple $(R_1, R_2, R_{Y,1}, R_{Y,1} + R_{Y,2})$ is said to be SR-OH achievable, if for any $\epsilon > 0$ and sufficiently large $n$, there exist an $(n, M_1, M_2, M_{Y,1}, M_{Y,2}, \epsilon, \epsilon)$

SR-OH code, such that

$$\frac{1}{n}\log M_1 \le R_1+\epsilon, \qquad \frac{1}{n}\log M_2 \le R_2+\epsilon,$$
$$\frac{1}{n}\log M_{Y,1} \le R_{Y,1}+\epsilon, \qquad \frac{1}{n}\log M_{Y,2} \le R_{Y,2}+\epsilon.$$

Denote the set of SR-OH achievable rate quadruples as $\mathcal{R}_{oh}$, and we seek to characterize this region for this problem. For easier comparison with the other problem, the last component of the rate vector is written as the sum-rate, instead of the individual rate $R_{Y,2}$. However it is straightforward to verify that $\mathcal{R}_{oh}$ is sufficient to provide a complete characterization if we were to define an achievable rate quadruple as the vector of $(R_1, R_2, R_{Y,1}, R_{Y,2})$.

For the single stage system, denote the minimum achievable rate at the $Y$ encoder for a given $X$ encoder rate $R$ as $R_{oh}(R)$, which is shown in [13], [14] to be

$$R_{oh}(R) = \min_{U}\{H(Y|U)|U \leftrightarrow X \leftrightarrow Y, I(X;U) \le R, |\mathcal{U}| \le |\mathcal{X}|+1\}. \quad (3)$$

From (2) and (3), it is clear that

$$R_{oh}(R) + E(R) = H(Y). \quad (4)$$

This suggests there is an intimate connection between the single stage hypothesis testing problem and the one-helper problem, and we shall explore this connection in the successive refinement coding case.

## III. Main Results

In the remainder of the work, for a given region $\mathcal{R}$ to be characterized, we shall use $\mathcal{R}^*$ to denote its single letter characterization form, and $\hat{\mathcal{R}}^*$ to denote its entropy characterization form. Our plan to characterize the regions $\mathcal{R}_{ht}$ and $\mathcal{R}_{oh}$ is as follows. First we provide an entropy characterization form of $\mathcal{R}_{ht}$, then give two equivalent forms of $\mathcal{R}_{oh}$: one is a single letter characterization while the other is in the entropy characterization form. Through the entropy characterization form, the SR-HT problem and SR-OH problem are shown to have intimate connection, by which a single letter characterization is established. Further connections between the problems, the new interpretation of the operational meaning of the information bottleneck method, and the relationship to the pattern recognition problem investigated in [19], [21] are subsequently discussed.

### A. Entropy characterization form of $\mathcal{R}_{ht}$

It is convenient to introduce the set $\mathcal{F}_n$ as the collection of functions with domain $\mathcal{X}^n$. First we define the following set

$$\hat{\mathcal{R}}^*_{ht} \triangleq \mathbf{CL}\bigcup_n \hat{\mathcal{R}}^*_{ht,n},$$

where

$$\hat{\mathcal{R}}^*_{ht,n} = \bigcup_{f_1,f_2\in\mathcal{F}_n}\Big\{(R_1,R_2,E_1,E_2): R_1 \ge \frac{1}{n}\log|f_1|, R_1+R_2 \ge \frac{1}{n}\log|f_1|+\frac{1}{n}\log|f_2|, E_1 \le \frac{1}{n}D(P_{f_1(X^n)Y^n}||P_{f_1(X^n)}P_{Y^n}), E_2 \le \frac{1}{n}D(P_{f_1(X^n)f_2(X^n)Y^n}||P_{f_1(X^n)f_2(X^n)}P_{Y^n})\Big\},$$

where

$$D(P||Q) \triangleq \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}$$

is the Kullback-Leibler information divergence. Note that $\bigcup_n \hat{\mathcal{R}}^*_{ht,n}$ is not necessarily a closed set, and thus we take its closure, denoted by $\mathbf{CL}$.

We can now follow the approach taken by Ahlswede and Csiszár and use Stein's lemma [28] to establish a relation between $\mathcal{R}_{ht}(\epsilon_1,\epsilon_2)$ and $\hat{\mathcal{R}}^*_{ht}$, which leads to a characterization of $\mathcal{R}_{ht}$ as a corollary.

***Theorem** 1:*

a) $\mathcal{R}_{ht}(\epsilon_1,\epsilon_2) \supseteq \hat{\mathcal{R}}^*_{ht}$, for all $\epsilon_1,\epsilon_2 \in (0,1)$.

b) $\mathcal{R}_{ht} \subseteq \hat{\mathcal{R}}^*_{ht}$.

This theorem is a generalization of the one given in [9], and the proof is thus omitted; interested readers can refer to [29] for more details. With Theorem 1 and the definition of $\mathcal{R}_{ht}$, it is straightforward to see that the following corollary is true.

***Corollary** 1:* $\mathcal{R}_{ht} = \hat{\mathcal{R}}^*_{ht}$.

Note further that

$$\frac{1}{n}D(P_{f_1(X^n)Y^n}||P_{f_1(X^n)}P_{Y^n}) = \frac{1}{n}I(f_1(X^n);Y^n) = H(Y) - \frac{1}{n}H(Y^n|f_1(X^n)),$$

and

$$\frac{1}{n}D(P_{f_1(X^n)f_2(X^n)Y^n}||P_{f_1(X^n)f_2(X^n)}P_{Y^n}) = \frac{1}{n}I(f_1(X^n)f_2(X^n);Y^n) = H(Y) - \frac{1}{n}H(Y^n|f_1(X^n)f_2(X^n)).$$

Thus it follows that

$$\hat{\mathcal{R}}^*_{ht,n} = \bigcup_{f_1,f_2\in\mathcal{F}_n}\Big\{(R_1,R_2,E_1,E_2): R_1 \ge \frac{1}{n}\log|f_1|, R_1+R_2 \ge \frac{1}{n}\log|f_1|+\frac{1}{n}\log|f_2|, H(Y)-E_1 \ge \frac{1}{n}H(Y^n|f_1(X^n)), H(Y)-E_2 \ge \frac{1}{n}H(H^n|f_1(X^n)f_2(X^n))\Big\}.$$

### B. Two equivalent characterizations of $\mathcal{R}_{oh}$

Next two equivalent characterizations of $\mathcal{R}_{oh}$ are given. One of them is in a single letter form, while the other is in the

entropy characterization form. Through the latter form, it will be clear there is an intimate connection between $\mathcal{R}_{oh}$ and $\mathcal{R}_{ht}$.

Define the region $\mathcal{R}^*_{oh}$ to be the set of all rate quadruples $(R_1, R_2, R_{Y,1}, R_{Y,1}+R_{Y,2})$ for which there exist random variables $(U, V)$ in finite alphabets $\mathcal{U}, \mathcal{V}$ such that the following conditions are satisfied.

1) $(U, V) \leftrightarrow X \leftrightarrow Y$ is a Markov string.
2) The non-negative rates $R_1$, $R_2$, $R_{Y,1}$ and $R_{Y,2}$ satisfy:

$$R_1 \geq I(X;U), \quad R_1 + R_2 \geq I(X;U,V),$$
$$R_{Y,1} \geq H(Y|U,V), \quad R_{Y,1} + R_{Y,2} \geq H(Y|U).$$

3) The alphabets $\mathcal{U}$, $\mathcal{V}$ satisfy

$$|\mathcal{U}| \leq |\mathcal{X}| + 3, \quad |\mathcal{V}| \leq |\mathcal{X}|^2 + 3|\mathcal{X}| + 1.$$

Note that the region $\mathcal{R}^*_{oh}$ is a closed set since entropy and mutual information are both continuous functions of each argument. We have the following theorem.

***Theorem** 2:* $\mathcal{R}_{oh} = \mathcal{R}^*_{oh}$.

In the proof of this theorem, we only outline the random coding argument for the achievability of the region; the converse is by generalizing the proof for the single stage case in [30], thus it is omitted (see [29] for details). It is worth pointing out that the achievability is proved by strategically combining the coding schemes for the original one-helper problem [13][14], the incremental Slepian-Wolf coding approach (see [31]–[34]), and the successive refinement source coding problem [2], [3].

*Proof:*

Let $\delta_i, i = 1, 2, 3$ be small positive quantities. Fix a probability distribution $P_{UVXY} = P_X P_{UV|X} P_{Y|X}$. First generate $2^{n(I(X;U)+\delta_1)}$ codewords single-letter-wise according to the distribution $P_U$, and denote the codebook as $\mathcal{C}_u$. For each of these $U$ codewords, generate $2^{n(I(X;V|U)+\delta_2)}$ codewords according to $P(V|U)$, and denote the codebook as $\mathcal{C}_v(u^n)$ for each $u^n \in \mathcal{C}_u$. This will be the codebook for the encoder observing source $X$. For the encoder observing $Y$, first construct a two-level nested binning structure, such that each coarser bin contains $2^{nI(Y;V|U)}$ smaller bins, with a total of $2^{n(H(Y|U,V)+\delta_3)}$ coarser bins; this induces a total of $2^{n(H(Y|U)+\delta_3)}$ finer bins. Assign each $y^n$ uniformly at random into one of finer bins. The codebooks are revealed to both the encoders and decoders.

During encoding, with high probability the encoder observing $x^n$ can find a codeword $u^n(i) \in \mathcal{C}_u$ that is jointly typical with $x^n$, and the index $i$ is sent to the decoder as the first stage description; for the given $u^n(i)$ codeword, again with high probability there exists a $v^n(j|i) \in \mathcal{C}_v(u^n(i))$ that is jointly typical with $x^n$ and $u^n(i)$. The index $j$ is sent as the second stage information. At the encoder observing $y^n$, the coarse bin index $k$ to which $y^n$ belongs is sent as the first stage information, while the finer bin index $l$ within the coarser bin is sent as the second stage information. The first decoder, with indices $i, k, l$ decodes, if it finds a unique $y^n$ sequence in the $(k, l)$-th finer bin that is jointly typical with $u^n(i)$; the second decoder, with indices $i, j, k$, decodes if it finds a unique $y^n$ sequence in the $k$-th coarser bin that is jointly typical with $u^n(i)$ and $v^n(j|i)$. Using a similar argument as for the original one-helper problem (see for example [30]), it can be shown that the above coding scheme succeeds with probability arbitrarily close to 1. ■

Next we give another characterization of $\mathcal{R}_{oh}$. Define the following set

$$\hat{\mathcal{R}}^*_{oh} = \mathbf{CL}\bigcup_n \hat{\mathcal{R}}^*_{oh,n},$$

where

$$\begin{aligned}\hat{\mathcal{R}}^*_{oh,n} = \bigcup_{f_1, f_2 \in \mathcal{F}_n} \{(R_1, R_2, R_{Y,1}, R_{Y,1}+R_{Y,2}) : \\ R_1 \geq \frac{1}{n}\log|f_1|, R_1 + R_2 \geq \frac{1}{n}\log|f_1| + \frac{1}{n}\log|f_2|, \\ R_{Y,1} \geq \frac{1}{n}H(Y^n|f_1(X^n)f_2(X^n)), \\ R_{Y,1} + R_{Y,2} \geq \frac{1}{n}H(Y^n|f_1(X^n))\Big\}.\end{aligned}$$

We have the following theorem.

***Theorem** 3:* $\mathcal{R}_{oh} = \hat{\mathcal{R}}^*_{oh} = \mathcal{R}^*_{oh}$.

*Proof:* To prove $\mathcal{R}_{oh} \supseteq \hat{\mathcal{R}}^*_{oh}$, we can either apply the (incremental) Slepian-Wolf coding scheme [31] on the super-source $Y^n$ with two degraded side information $f_1(X^n)$ and $(f_1(X^n)f_2(X^n))$, or apply Heegard-Berger coding theorem [33] on the super-source; note here $(f_1(X^n), f_2(X^n), Y^n)$ are i.i.d. random variables across blocks. The details are omitted.

To see $\mathcal{R}_{oh} \subseteq \hat{\mathcal{R}}^*_{oh}$, we write

$$\begin{aligned}nR_{Y,1} &\geq H(f_{Y,1}(Y^n)) \geq H(f_{Y,1}(Y^n)|f_1(X^n)f_2(X^n)) \\ &= I(Y^n; f_{Y,1}(Y^n)|f_1(X^n)f_2(X^n)) \\ &\overset{(a)}{\geq} H(Y^n|f_1(X^n)f_2(X^n)) - n\log|\mathcal{Y}|\Delta_1 - H_b(\Delta_1)\end{aligned}$$

where (a) is by applying Fano's inequality. The other condition on the sum rate can be proved similarly. ■

### C. Connection between the SR-HT and SR-OH problems

Define the *partially skewed reflection* operator as follows.

***Definition** 6:* For a real quadruple $(a_1, a_2, a_3, a_4)$, its partially skewed reflection operation under a two-source $(\mathcal{X}, \mathcal{Y}, P_{XY})$ is given by

$$\mathbf{P}(a_1, a_2, a_3, a_4) = (a_1, a_2, H(Y) - a_4, H(Y) - a_3).$$

The partially skewed reflection of a set of quadruples $S$ is given by

$$\mathbf{P}(S) = \{\mathbf{P}(a_1, a_2, a_3, a_4) : (a_1, a_2, a_3, a_4) \in S\}.$$

Since $\mathbf{P}$ is clearly a bijection and preserves Euclidean distance, it is an isometry. From Theorem 1, Theorem 3 and the corresponding entropy characterization expressions, we have the following corollary.

***Corollary** 2:* $\mathcal{R}_{ht} = \mathbf{P}(\mathcal{R}_{oh})$.

The isometry $\mathbf{P}$ implies the two regions are congruent. Since $\mathcal{R}_{oh}$ is convex, $\mathcal{R}_{ht}$ is also convex. This fact does not directly follow from the time-sharing argument as often seen in source coding, because the time-sharing argument does not directly apply in the SR-HT problem. Furthermore, since a single letter characterization of $\mathcal{R}_{oh}$ is available, we thus readily find

a single-letter characterization for $\mathcal{R}_{ht}$; for convenience we denote it as $\mathcal{R}^*_{ht}$.

The connection among the two problems can be further strengthened. For an arbitrary point on the boundary of $\mathcal{R}_{ht}$, by the isometry of $\mathbf{P}$, there is one point on the boundary of $\mathcal{R}_{oh}$. Using the entropy characterization form, it is clear that there exists an optimal sequence of functions $f_1, f_2 \in \mathcal{F}_n$, in the sense that the values of

$$(\frac{1}{n}\log|f_1|, \frac{1}{n}\log|f_2|, \frac{1}{n}I(Y^n; f_1(X^n)), \frac{1}{n}I(Y^n; f_1(X^n)f_2(X^n)))$$

approach the particular operating points for SR-HT, as well as the corresponding point for SR-OH problem. Denote the concatenation of these functions with $m$ such $n$-blocks as $f_1^m$ and $f_2^m$. It is seen that when $m$ is sufficiently large (with the sequence of code $f_1, f_2 \in \mathcal{F}_n$), the sequence of codes $(f_1^m, f_2^m) \in \mathcal{F}_{mn}$ is indeed approaching optimum. Thus we have the following theorem.

***Theorem** 4:* For any particular point $\boldsymbol{s}$ in $\mathcal{R}_{ht}$ and the corresponding point $\mathbf{P}(\boldsymbol{s}) \in \mathcal{R}_{oh}$, there exists a sequence of optimal coding functions $(f_1, f_2) \in \mathcal{F}_{l_n}$, where $l_n \to \infty$ as $n \to \infty$, in the sense that they approach $\boldsymbol{s}$, and there exists a corresponding sequence of coding functions $f_{Y,1}, f_{Y,2} \in \mathcal{F}_{l_n}$, such that the sequence of these four coding functions approaches the point $\mathbf{P}(\boldsymbol{s}) \in \mathcal{R}_{oh}$.

It is now clear that the two problems are closely related and can be treated together. In Section IV, we consider the notion of successive refinability in the two settings together, and derive necessary and sufficient conditions; a binary source example will also be considered in this context.

### *D. The strong converse result for the SR-HT problem*

Though $\mathcal{R}_{ht}$ can be characterized in a single-letter form as above, this is not sufficient to characterize $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$ with arbitrary $\epsilon_1, \epsilon_2 \in (0,1)$. As it turns out $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$ is almost independent of $(\epsilon_1, \epsilon_2)$. We have the following strengthened result, the converse part of which is proved in Section VI using the method of types.

***Theorem** 5:* For any $\epsilon_1, \epsilon_2 \in (0,1)$ such that $\epsilon_1 + \epsilon_2 < 1$, $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2) = \mathcal{R}_{ht}$. On the other hand, for any $\epsilon_1, \epsilon_2 \in (0,1)$ such that $\epsilon_1 + \epsilon_2 > 1$, we have

$$\mathcal{R}_{ht}(\epsilon_1, \epsilon_2) = \{(R_1, R_2, E_1, E_2) : E_1 \le E(R_1), E_2 \le E(R_1 + R_2)\}. \quad (5)$$

Note the case $\epsilon_1 + \epsilon_2 = 1$ is not included. This is similar to the source-channel separation results when the entropy rate is exactly equal to the channel capacity, the behavior is not known. The achievability result for the case $\epsilon_1 + \epsilon_2 < 1$ is implied by Theorem 1, and next we give the achievability proof for the other case.

*Proof of achievability for Theorem 5:*

Since $E(R_1)$ and $E(R_1 + R_2)$ are achievable type-two error exponents with coding rate $R_1$ and $R_1 + R_2$ for single stage coding, respectively, it follows that there exist encoding functions $f_1'$ and $f_2'$, and the corresponding detectors $g'_{t,1}$ and $g'_{t,2}$ to approach this performance. Denote the acceptance regions as $A_1'$ and $A_2'$, and type-two errors by $f_1'$ and $f_2'$ as $\beta_1'$ and $\beta_2'$, respectively; note that the type-one errors $\epsilon_1'$ and $\epsilon_2'$ can be made arbitrarily small when $n$ is sufficiently large.

We now construct a two-stage system using these functions. Given fixed $\epsilon_1$ and $\epsilon_2$ such that $\epsilon_1 + \epsilon_2 > 1$, we partition the $\mathcal{X}^n$ space into two non-intersecting sets $A$ and $B$, such that $P_X^n(A) > 1 - \epsilon_1$ and $P_X^n(B) > 1 - \epsilon_2$; with sufficiently large $n$ such a partition is always possible. Note that $P_X^n(A) + P_X^n(B) = 1$. The encoding is performed as follows. In the first stage, if $\boldsymbol{x} \in A$, then $f_1'$ is used; if $\boldsymbol{x} \in B$, then send the first $nR_1$ bits of $f_2'$. In the second stage, if $\boldsymbol{x} \in A$, we send a fixed codeword of length $nR_2$; if $\boldsymbol{x} \in B$, then we send the remaining $nR_2$ bits of $f_2'$. An additional prefix bit is added to indicate which set $\boldsymbol{x}$ is in, and this induces a negligible rate increase for long block codes. With this prefix bit, the first stage decoder uses the following decision set, which indeed utilizes only $n(R_1)$ bits (plus the one prefix bit) of the description

$$C_1 = (A \times \mathcal{Y}^n) \cap A_1'.$$

For the second stage detector, the following decision region is used

$$C_2 = (B \times \mathcal{Y}^n) \cap A_2'.$$

It remains to show the error probabilities are as claimed. Note that $P_{f_1'(X^n)Y^n}(A_1') \ge 1 - \epsilon_1'$ and the inequality $P_X^n(A) > 1 - \epsilon_1$ is strict, thus by applying the union bound

$$P_{f_1(X^n)Y^n}(C_1) \ge P_X^n(A) - \epsilon_1' \ge 1 - \epsilon_1,$$

when $n$ is sufficiently large, i.e., $\epsilon_1'$ is sufficiently small. Similarly for the second stage

$$P^n_{f_1(X^n)f_2(X^n)Y^n}(C_2) \ge P_X^n(B) - \epsilon_2' \ge 1 - \epsilon_2,$$

when $n$ is sufficiently large since $P_X^n(B) > 1 - \epsilon_2$. For the type-two errors, we have

$$\begin{aligned} P_{f_1(X^n)}P_Y^n(C_1) &\le P_{f_1'(X^n)}P_Y^n(A_1') = \beta_1' \\ P_{f_1(X^n)f_2(X^n)}P_Y^n(C_2) &\le P_{f_1'(X^n)f_2'(X^n)}P_Y^n(A_2') = \beta_2'. \end{aligned}$$

This indeed implies the claimed result and the proof is complete. ■

### *E. Connection to the pattern recognition problem*

The successive refinement pattern recognition (SR-PR) problem was formulated independently by Tuncel [19] and by Westover and O'Sullivan [23]. In this setting, a two-source $(\mathcal{X}, \mathcal{Y}, P_{XY})$ is an *environment* $\mathcal{E}$ for a pattern recognition system. The pattern domain is $\mathcal{X}^n$ and the noisy observation domain is $\mathcal{Y}^n$. We provide a brief problem definition below, and more details can be found in [19]–[23].

***Definition** 7:* An $(n, M_{c,1}, M_{c,2})$ instance of the environment $\mathcal{E}$ consists of $M_{c,1}M_{c,2}$ $n$-length sequences in $\mathcal{X}^n$, labeled as $X^n(1), X^n(2), ..., X^n(M_{c,1}M_{c,2})$.

***Definition** 8:* An $(n, M_{c,1}, M_{c,2}, M_1, M_2, \Delta_1, \Delta_2)$ SR-PR code for an environment $\mathcal{E}$ consists of two encoders

$$f_1 : \mathcal{X}^n \to I_{M_1}, \quad f_2 : \mathcal{X}^n \to I_{M_2},$$

and two classifiers

$$g_{r,1} : I_{M_1}^{M_{c,1}} \times \mathcal{Y}^n \to I_{M_{c,1}},$$
$$g_{r,2} : I_{M_1}^{M_{c,1}M_{c,2}} \times I_{M_2}^{M_{c,1}M_{c,2}} \times \mathcal{Y}^n \to I_{M_{c,1}M_{c,2}}.$$

We denote $J_1(m) = f_1(X^n(m))$ and $J_2(m) = f_2(X^n(m))$; furthermore denote the collection of codewords as $\mathcal{C}_1$ and $\mathcal{C}_2$ for an $(n, M_{c,1}, M_{c,2})$ instance of the environment, i.e.,

$$\mathcal{C}_1 = \{J_1(1), J_1(2), ..., J_1(M_c)\}$$
$$\mathcal{C}_2 = \{[J_1(1), J_2(1)], [J_1(2), J_2(2)], ..., [J_1(M_{c,1}M_{c,2}), J_2(M_{c,1}M_{c,2})]\}.$$

In the recognition phase of the system, the pattern occurs uniformly at random in the pattern pool given in the enrollment phase. More precisely, a random pattern $W_1 \in I_{M_{c,1}}$ occurs either uniformly at random in the $M_{c,1}$ given patterns where the first level description will be used, or a random pattern $W_2 \in I_{M_{c,1}M_{c,2}}$ occurs uniformly at random in the $M_{c,1}M_{c,2}$ given patterns where both levels of descriptions will be used. For a given system, the error probability for the first level $P_{e,1}$ and that for the second level $P_{e,2}$ satisfy, respectively,

$$\Delta_1 \geq P_{e,1} \triangleq \Pr\{g_{r,1}(\mathcal{C}_1, Y^n(W_1)) \neq W_1\},$$
$$\Delta_2 \geq P_{e,2} \triangleq \Pr\{g_{r,2}(\mathcal{C}_2, Y^n(W_2)) \neq W_2\}.$$

Note that both $(\mathcal{C}_1, \mathcal{C}_2)$ and $(W_1, W_2)$ are random quantities.

***Definition** 9:* A rate vector $(R_1, R_2, R_{c,1}, R_{c,1} + R_{c,2})$ is SR-PR achievable, if for any $\epsilon > 0$ and sufficiently large $n$ there exists an $(n, M_{c,1}, M_{c,2}, M_1, M_2, \epsilon, \epsilon)$ code such that

$$\frac{1}{n}\log M_{c,1} \geq R_{c,1} - \epsilon, \qquad \frac{1}{n}\log M_{c,2} \geq R_{c,2} - \epsilon$$
$$\frac{1}{n}\log M_1 \leq R_1 + \epsilon, \qquad \frac{1}{n}\log M_2 \leq R_2 + \epsilon.$$

Denote the set of achievable rate quadruples for SR-PR as $\mathcal{R}_{pr}$, and a characterization was given in [19][23]. By comparing the expression provided there, it is not difficult to see $\mathcal{R}_{pr} = \mathcal{R}_{ht}$. In fact, the entropy characterization approach given in this work provides a simple alternative proof for the pattern recognition rate region.

### *F. An interpretation of the information bottleneck method*

The information bottleneck function was given in [24] as

$$R_{ib}(R) = \min_{I(U;Y)\geq R,\, U\leftrightarrow X\leftrightarrow Y} I(U;X), \tag{6}$$

which is exactly the definition of the inverse function of $E(R)$ in (2) if we ignore the cardinality bound. This similarity motivates the following definition of an information bottleneck code, extended to its successive refinement version.

***Definition** 10:* An $(n, M_1, M_2, R_{I,1}, R_{I,2}, \Delta_1, \Delta_2)$ SR-IB code for source $(X, Y)$ consists of two classification functions

$$f_1 : \mathcal{X}^n \to I_{M_1}, \qquad f_2 : \mathcal{X}^n \to I_{M_2},$$

such that

$$\frac{1}{n}I(Y^n; f_1(X^n)) \geq R_{I,1} + \Delta_1,$$
$$\frac{1}{n}I(Y^n; f_1(X^n)f_2(X^n)) \geq R_{I,2} + \Delta_2.$$

***Definition** 11:* A rate quadruple $(R_1, R_2, R_{I,1}, R_{I,2})$ is said to be SR-IB achievable, if for any $\epsilon > 0$ and sufficiently large $n$, there exists an $(n, M_1, M_2, R_{I,1}, R_{I,2}, \epsilon, \epsilon)$ SR-IB code, such that

$$\frac{1}{n}\log M_1 \leq R_1 + \epsilon, \qquad \frac{1}{n}\log M_2 \leq R_2 + \epsilon.$$

Denote the set of achievable rate quadruples for SR-IB as $\mathcal{R}_{ib}$, and thus this is the region of interest. The following theorem is immediate.

***Theorem** 6:* $\mathcal{R}_{ib} = \mathcal{R}_{pr} = \mathcal{R}_{ht}$.

*Proof:* We only need to show that $\hat{\mathcal{R}}_{ht}^* = \mathcal{R}_{ib}$. The inclusion $\mathcal{R}_{ib} \subseteq \hat{\mathcal{R}}_{ht}^*$ is rather trivial by the definitions. For the inclusion in the other direction, observe that for any fixed-rate SR-IB code of length $k$, by taking its $l$-fold product codes, we can easily show that

$$I(f_1(X_1^n), f_1(x_{n+1}^{2n}), ..., f_1(x_{n(l-1)+1}^{ln}); Y_1^{ln}) = lI(f_1(X_1^n); Y_1^n)),$$

Thus we have $\hat{\mathcal{R}}_{ht}^* \subseteq \mathcal{R}_{ib}$, which establishes $\hat{\mathcal{R}}_{ht}^* = \mathcal{R}_{ib}$. ■

The above formalization of the operational meaning of the IB function essentially states that we can understand the IB problem as a source coding problem subject to a constraint on the normalized mutual information between the codeword and the remote source vector $Y^n$, instead of the usual single-letter distortion measure familiar in the rate-distortion theory. Moreover, the IB problem is not uncommon in multi-terminal systems, though it might appear in certain disguise, as shown by the problems in consideration.

## IV. Successive refinability in SR-HT and SR-OH

### *A. Successive refinability*

Similar to the notion of successive refinability in the rate-distortion setting, we can introduce the following notions for the two problems considered in this work. These notions capture whether the progressive coding requirement causes loss of performance with respect to single-stage coding.

***Definition** 12:* A source is successively refinable for hypothesis testing (with $(\epsilon_1, \epsilon_2)$) and one-helper coding, respectively, with rate $R_1$ and $R_2$ if

$$(R_1, R_2, E(R_1), E(R_1 + R_2)) \in \mathcal{R}_{ht}(\epsilon_1, \epsilon_2),$$
$$(R_1, R_2, R_{oh}(R_1 + R_2), R_{oh}(R_1)) \in \mathcal{R}_{oh}.$$

Note we can also define *weakly successive refinability* for SR-HT as $(R_1, R_2, E(R_1), E(R_1 + R_2)) \in \mathcal{R}_{ht}$. This weaker notion will be useful when Gaussian source is considered, for which Theorem 5 does not apply because of its reliance on the method of types. Using the characterization of $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$, we have the following theorem for the SR-HT problem.

***Theorem** 7:* 1) If $\epsilon_1, \epsilon_2 \in (0, 1)$ such that $\epsilon_1 + \epsilon_2 < 1$, a two-source $(\mathcal{X}, \mathcal{Y}, P_{XY})$ is successively refinable for SR-HT with rate $R_1$ and $R_2$, if and only if there exist random variables $U$ and $V$ in finite alphabets $\mathcal{U}$ and $\mathcal{V}$ such that

a) $Y \leftrightarrow X \leftrightarrow V \leftrightarrow U$ is a Markov string.
b) $I(X;U) = R_1$ and $I(Y;U) = E(R_1)$.
c) $I(X;V) = R_1 + R_2$ and $I(Y;V) = E(R_1 + R_2)$.

2) If $\epsilon_1, \epsilon_2 \in (0,1)$ such that $\epsilon_1 + \epsilon_2 > 1$, a two-source $(\mathcal{X}, \mathcal{Y}, P_{XY})$ is always successively refinable for SR-HT with rate $R_1$ and $R_2$.

*Proof:* Note that part (2) follows directly from Theorem 5, and thus we only consider part (1). Because of the relation $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2) = \mathcal{R}_{ht} = \mathcal{R}^*_{ht}$ for this case, $\mathcal{R}^*_{ht}$ is sufficient to characterize the region. Note that in the definition of $\mathcal{R}^*_{ht}$ we can always add in the Markov string condition $X \leftrightarrow V \leftrightarrow U$ by letting $V = (U, V)$, which does not change any involved information quantities. This necessitates increasing the cardinality bound of $\mathcal{V}$, and it is trivial to see that a size of $(|\mathcal{X}| + 3)(|\mathcal{X}|^2 + 3|\mathcal{X}| + 1)$ suffices. This observation alone provides the following alternative definition of $\mathcal{R}^*_{ht}$ as the set of quadruples $(R_1, R_2, E_1, E_2)$ for which there exist random variables $(U, V)$ in finite alphabets $\mathcal{U}, \mathcal{V}$ such that:

1) $Y \leftrightarrow X \leftrightarrow V \leftrightarrow U$ is a Markov string.
2) The non-negative rate quadruple satisfies:

$$R_1 \geq I(X;U), \qquad R_1 + R_2 \geq I(X;V),$$
$$E_1 \leq I(Y;U), \qquad E_2 \leq I(Y;V).$$

3) $|\mathcal{U}| \leq |\mathcal{X}| + 3$, and $|\mathcal{V}| \leq (|\mathcal{X}| + 3)(|\mathcal{X}|^2 + 3|\mathcal{X}| + 1)$.

Now the necessity and sufficiency both follow directly from this characterization. ∎

The results can clearly be extended to SR-OH with virtually no change (without the second part); we thus omit the statement of such a theorem.

### B. The doubly symmetric binary source

Consider the following hypothesis: $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $0 \leq p_0 < 0.5$,

$$H_0 : P_{XY}(x,y) = \frac{1}{2}(1-p_0)\delta_{x,y} + \frac{1}{2}p_0(1-\delta_{x,y}),$$
$$H_1 : P_X(x)P_Y(y) = \frac{1}{4}.$$

For $H_0$, the probability distribution $P_{XY}$ can essentially be understood as there is a binary symmetric channel (BSC) with crossover probability $p_0$ with input $X$ and output $Y$, and the input $X$ is of distribution Bernoulli $\frac{1}{2}$ (denoted as Bern($\frac{1}{2}$)).

In [18], Wyner showed that the optimal forward test channel for the single stage one-helper problem is given by $U = X \oplus N$, where $\oplus$ is modulo 2 addition and $N$ is a Bern $(H_b^{-1}(1-R))$ random variable, independent of everything else; here $H_b^{-1}(\cdot)$ denotes the inverse of the binary entropy function $H_b(p)$ with $p \in [0, 0.5]$, and $R$ is the coding rate at encoder observing $X$. It is seen that when successive refinement coding is used, we can choose $U = X \oplus N_1 \oplus N_2$ and $V = X \oplus N_1$, where $N_1$ is of Bern$(H_b^{-1}(1-R_1-R_2))$ and $N_2$ is a Bernoulli random variable such that $N_1 + N_2$ is of Bern$(H_b^{-1}(1-R_1))$; such an $N_2$ always exists since $H_b^{-1}(1-R_1) \geq H_b^{-1}(1-R_1-R_2)$. $N_1$ and $N_2$ are independent of each other and everything else. By the optimality of this forward test channel shown in [18], $(U, V)$ clearly satisfies the conditions in Theorem 7, and thus for the SR-HT problem (as well as successive refinement pattern recognition problem and the information bottleneck problem), it is indeed successively refinable with any rate $R_1$ and $R_2$.

This example highlights the power of treating these problems together. In [19] the same result was given for the pattern recognition problem, and the derivation is rather non-trivial. By recognizing the relation among these problems, we simply invoke the existing result in [18] to avoid such difficulty.

## V. The Gaussian Source

Until this point, we have only considered discrete memoryless sources. The results however can be extended to more general source such as the Gaussian source. It is not difficult to verify that the converse proof for the SR-HT problem can be established using the almost identical line of derivation as in the SR-OH problem by bounding $I(Y^n; f_1(X^n))$ and $I(Y^n; f_1(X^n)f_2(X^n))$ directly. Next we provide an achievability proof using a lattice strategy for the SR-HT problem; one can also invoke the result on the pattern recognition problem directly to obtain such a proof, however the method below is more constructive. For $H_0$, let the distribution $P_{XY}$ be given as $Y = X + N$, where $X \sim \mathcal{N}(0, \sigma_x^2)$ and $N \sim \mathcal{N}(0, \sigma_N^2)$ are independent; for $H_1$, $X$ and $Y$ are independent with the distributions given by the marginal distribution of $P_{XY}$.

Before considering the lattice strategy, let us derive an explicit outer bound for $\mathcal{R}_{ht}$. We have that

$$\begin{aligned} I(U;Y) &= h(Y) - h(X+N|U) \\ &\overset{(a)}{\leq} h(Y) - \frac{1}{2}\log[2\pi e\sigma_N^2 + \exp(2h(X|U))] \\ &= h(Y) - \frac{1}{2}\log[2\pi e\sigma_N^2 + \exp(2h(X) - 2I(X;U))] \\ &\overset{(b)}{\leq} h(Y) - \frac{1}{2}\log[2\pi e\sigma_N^2 + 2\pi e\sigma_x^2\exp(-2R_1)] \\ &= \frac{1}{2}\log\frac{\sigma_x^2+\sigma_N^2}{\sigma_N^2+\sigma_x^2\exp(-2R_1)}, \end{aligned}$$

where (a) is by applying the conditional form of the entropy power inequality [35] and (b) is because $I(X;U) \leq R_1$. Similarly we have

$$I(UV;Y) \leq \frac{1}{2}\log\frac{\sigma_x^2+\sigma_N^2}{\sigma_N^2+\sigma_x^2\exp[-2(R_1+R_2)]}.$$

The construction relies on the entropy-coded dithered quantization (ECDQ), the details of which can be found in [36]–[38]. An $n$-dimensional lattice quantizer is formed by a lattice $\Lambda_n$. The quantizer $Q_n(\cdot)$ maps each vector $\boldsymbol{x} \in \Re^n$ into the lattice point $\boldsymbol{\lambda}_i \in \Lambda_n$ that is nearest to $\mathbf{x}$. The region of all $n$-vectors mapped into a lattice point $\boldsymbol{\lambda}_i \in \Lambda_n$ is the Voronoi region

$$V(\boldsymbol{\lambda}_i) = \{\boldsymbol{x} \in \Re^n : ||\boldsymbol{x} - \boldsymbol{\lambda}_i|| \leq ||\boldsymbol{x} - \boldsymbol{\lambda}_j||, \forall j \neq i\}.$$

The dither $\boldsymbol{Z}$ is an $n$-dimensional random vector, independent of the source, and uniformly distributed over the basic cell $V_0$ of the lattice which is the Voronoi region of the lattice point $\mathbf{0}$. The dither vector is assumed to be available to both the encoder and the decoder. The normalized second moment $G_n$ of the lattice characterizes the second moment of the dither vector

$$\frac{1}{n}\mathbb{E}||\boldsymbol{Z}||^2 = G_n V^{2/n},$$

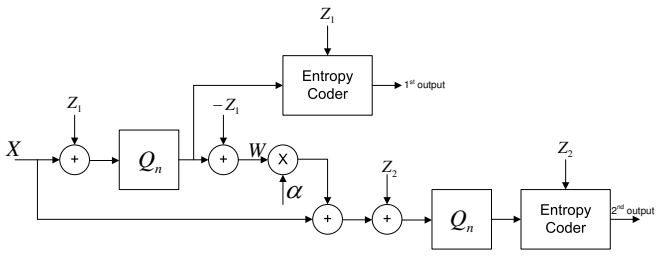


Fig. 3. Encoder based on ECDQs.

where $V$ denotes the volume of $V_0$. Both the entropy encoder and the decoder are conditioned on the dither sample $\mathbf{Z}$; furthermore, the entropy coder is assumed to be ideal. The dithered lattice quantizer represents the source vector $\mathbf{X}$ by the vector $\mathbf{W} = Q_n(\mathbf{X} + \mathbf{Z}) - \mathbf{Z}$.

Now we describe the coding system using ECDQs, which is essentially a two-stage quantization system, with the additional detectors at the decoder. Note that instead of the distortion of each length-$n$ block, we are interested in the detection performance using multiple such length-$n$ blocks. The system consists of two stages. The first stage takes input $\boldsymbol{X}$ and passes it through an ECDQ module. The output $\boldsymbol{W} = Q_n(\boldsymbol{X} + \boldsymbol{Z_1}) - \boldsymbol{Z_1}$ is scaled by $\alpha = \frac{-\sigma_x^2}{\sigma_x^2+\sigma_1^2}$ and added with $\boldsymbol{X}$. The resulting vector $\boldsymbol{X} + \alpha\boldsymbol{W}$ is passed through another ECDQ whose output is given as $Q_n(\boldsymbol{X} + \alpha\boldsymbol{W} + \boldsymbol{Z_2}) - \boldsymbol{Z_2}$, where $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are independent. Note here we slightly abuse the notations by allowing $Q_n$ to be a lattice quantizer scaled by different constant, which are reflected by the variance of $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$, denoted as $\sigma_1^2$ and $\sigma_2^2$, respectively. The system is depicted in Fig. 3. The detectors do not make a decision on one $\boldsymbol{X}$ block of length-$n$, but do so after receiving many such blocks.

Under the reconstruction using ECDQ, the output is distributed as $\boldsymbol{X} + \boldsymbol{N_1}$, and the noise vector $\boldsymbol{N_1}$ is distributed uniformly over the basic cell of $\Lambda_n$. If $\boldsymbol{N_1}$ was a Gaussian vector, we would be able to explicitly derive the Neyman-Pearson detector, and analyze its performance. Though this is not the case, the lattice quantization noise is nevertheless quite close to Gaussian for high-dimensional quantizers, thus it is likely the Neyman-Pearson detector derived assuming Gaussian distribution will provide near optimal performance, which turns out to be indeed the case. Next we use large deviation method to analyze the performance of such an approximation. Some necessary notations and results from [39], [40] are reviewed first. For simplicity the single stage case is investigated first, after which the generalization to the two-stage case is straightforward.

For a lattice $\Lambda_n$, the covering radius $R_u$ is the radius of the smallest $n$-dimensional ball to cover the Voronoi region $V_0$. The effective radius $R_l$ is the radius of a sphere having the same volume as $V_0$. We will need the following quantity

$$\epsilon(\Lambda_n) \triangleq \log\left(\frac{R_u}{R_l}\right) + \frac{1}{2}\log 2\pi e G_n^* + \frac{1}{n},$$

where $G_n^*$ is the normalized second moment of an $n$-sphere. It was shown by Rogers [41], [42] that there exist lattices which satisfy

$$\log\frac{R_u}{R_l} \to 0,$$

as $n \to \infty$. This implies that for such lattices $\epsilon(\Lambda_n) \to 0$ as $n \to \infty$.

Denote by $\mathcal{B}(R_u)$ a ball of radius $R_u$ and let $\sigma_u^2$ be the second moment per dimension of $\mathcal{B}(R_u)$; denote the variance of $\boldsymbol{N_1}$ per dimension as $\sigma_1^2$. The following lemma was proved in [39] (Lemma 6 and Lemma 11).

***Lemma** 1:* Let $\boldsymbol{G}_1 \sim \mathcal{N}(0, \sigma_u^2 \cdot I^n)$, then for Rogers-good lattices, the density of the noise distribution $p_{\boldsymbol{N_1}}$ and $p_{\boldsymbol{G_1}}$ satisfy

$$\frac{1}{n}\log\frac{p_{\boldsymbol{N_1}}(\boldsymbol{x})}{p_{\boldsymbol{G_1}}(\boldsymbol{x})} \le \epsilon(\Lambda_n). \tag{7}$$

Furthermore we have

$$\frac{n+2}{n}\sigma_u^2 \ge \sigma_1^2 \ge \left(\frac{R_u}{R_l}\right)^2 \sigma_u^2. \tag{8}$$

This lemma implies that the probability density of $\boldsymbol{N}_1$ can approximately be upper bounded by a Gaussian distribution, whose variance is almost the same as that of the quantization noise $\sigma_1^2$ when $n$ is sufficiently large.

Let us assume $\boldsymbol{N}_1$ indeed has an independent Gaussian distribution and derive the Neyman-Pearson detector under this assumption. We have the likelihood ratio for length-$n$ sequences

$$\frac{p(\boldsymbol{X}+\boldsymbol{N_1}, \boldsymbol{Y})}{p(\boldsymbol{X}+\boldsymbol{N_1})p(\boldsymbol{Y})} = \frac{p(\boldsymbol{Y}|\boldsymbol{X}+\boldsymbol{N_1})}{p(\boldsymbol{Y})}$$
$$= \exp\left(-\frac{\|\boldsymbol{Y} - \frac{\sigma_x^2}{\sigma_x^2+\sigma_1^2}(\boldsymbol{X}+\boldsymbol{Z_1})\|^2}{2(\sigma_N^2 + \frac{\sigma_x^2\sigma_1^2}{\sigma_x^2+\sigma_1^2})}\right)\exp\left(\frac{\|\boldsymbol{Y}\|^2}{2(\sigma_N^2+\sigma_x^2)}\right).$$

Thus the Neyman-Pearson detector makes decision by thresholding the following quantity

$$T = (\sigma_N^2+\sigma_x^2)\left\|\boldsymbol{Y} - \frac{\sigma_x^2}{\sigma_x^2+\sigma_1^2}(\boldsymbol{X}+\boldsymbol{Z_1})\right\|^2 - (\sigma_N^2 + \frac{\sigma_x^2\sigma_1^2}{\sigma_x^2+\sigma_1^2})\|\boldsymbol{Y}\|^2.$$

This quantity is essentially a weighted distance difference in the Euclidean space. It is straightforward to verify that the expectation of this quantity under the two hypotheses is given by

$$\mathbb{E}(T|H_0) = 0, \quad \mathbb{E}(T|H_1) = 2n\frac{(\sigma_x^2+\sigma_N^2)\sigma_x^4}{\sigma_x^2+\sigma_1^2}.$$

Now we take $m$ blocks of $n$-dimensional ECDQ, and consider a length-$mn$ source block. Choose the threshold as $mn\delta$, where $\delta$ is a small positive quantity the meaning of which will be clear later: if $T \le mn\delta$, hypothesis $H_0$ is accepted. To bound the type-two error exponent, define the

following new random variable

$$T' = (\sigma_N^2 + \sigma_x^2) \left\| \boldsymbol{Y} - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_1^2}(\boldsymbol{X} + \boldsymbol{G_1}) \right\|^2 - (\sigma_N^2 + \frac{\sigma_x^2 \sigma_1^2}{\sigma_x^2 + \sigma_1^2}) \left\| \boldsymbol{Y} \right\|^2 .$$

Using the Gaussian distribution approximation in Lemma 1 gives

$$\beta_1 = \Pr(T \le mn\delta | H_1) \le \Pr(T' \le mn\delta | H_1) \exp(mn\epsilon(\Lambda_n)),$$

which is straightforwardly seen because $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{N_1}$ are mutually independent, and so are $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{G_1}$ under hypothesis $H_1$; furthermore the bound given in Lemma 1 is uniform.

The moment generating function of $T'$ can be computed as

$$\begin{aligned}
&\lambda_{T'}(t) = \mathbb{E}[\exp(tT')] \\
&= \mathbb{E}_{\boldsymbol{Y}} \left[ \mathbb{E}_{\boldsymbol{X}, \boldsymbol{G_1}} [\exp(tT') | \boldsymbol{Y}] \right] \\
&\overset{(a)}{=} (1 - 2ct)^{-\frac{mn}{2}} \mathbb{E}_{\boldsymbol{Y}} \left[ \exp(-b\|\boldsymbol{Y}\|^2) \exp(\frac{a\|\boldsymbol{Y}\|^2}{1 - 2ct}) \right] \\
&\overset{(b)}{=} (1 - 2ct - 2a^2 t + 2abt - 4abct^2)^{-mn/2}
\end{aligned}$$

where we have defined

$$a \triangleq \sigma_x^2 + \sigma_N^2, \quad b \triangleq \sigma_N^2 + \frac{\sigma_x^2 \sigma_1^2}{\sigma_x^2 + \sigma_1^2},$$
$$c \triangleq \frac{(\sigma_x^2 + \sigma_N^2)(\sigma_x^2 + \sigma_u^2)\sigma_x^4}{(\sigma_x^2 + \sigma_1^2)^2},$$

and (a) is true because conditioned on $\boldsymbol{Y}$, $T'$ has a non-central Chi-square distribution; (b) is true by recognizing again the Chi-square distribution. The moment generating function exists whenever

$$1 - 2ct > 0, \text{ and } 1 - 2ct - 2a^2 t + 2abt - 4abct^2 > 0. \quad (9)$$

By applying the Chernoff bound for $t \le 0$, it follows that

$$\Pr(T' \le mn\delta | H_1) \quad \le \quad \exp(-mnt\delta + \lambda_{T'}(t)).$$

This implies the error exponent satisfies

$$E_1 \ge t\delta + \frac{1}{2} \log(1 - 2ct - 2a^2 t + 2abt - 4abct^2) - \epsilon(\Lambda_n). \quad (10)$$

Optimizing over $t$ to maximize the second term in (10), we have

$$E_1 \ge t^*\delta - \epsilon(\Lambda_n) + \frac{1}{2} \log \left( \frac{\sigma_x^4 (2\sigma_x^2 + \sigma_1^2 + \sigma_u^2)^2}{4(\sigma_x^2 + \sigma_1^2)(\sigma_x^2 + \sigma_u^2)(\sigma_x^2 \sigma_1^2 + \sigma_x^2 \sigma_N^2 + \sigma_N^2 \sigma_1^2)} + 1 \right) \quad (11)$$

where

$$t^* = \frac{-(2\sigma_x^2 + \sigma_1^2 + \sigma_u^2)(\sigma_x^2 + \sigma_1^2)}{4(\sigma_x^2 + \sigma_N^2)(\sigma_x^2 + \sigma_u^2)(\sigma_x^2 \sigma_1^2 + \sigma_x^2 \sigma_N^2 + \sigma_N^2 \sigma_1^2)}.$$

Define the right-hand-side of (11) as $E_1^*$. It can be easily checked that both the conditions in (9) are satisfied. We can choose $\delta$ sufficiently small, as long as it is positive; furthermore, by Lemma 1, and by making $n$ sufficiently large, $\epsilon(\Lambda_n)$ can be made arbitrarily small, and $\sigma_u^2 \to \sigma_1^2$, thus we have

$$E_1^* \to \frac{1}{2} \log \frac{(\sigma_x^2 + \sigma_N^2)(\sigma_x^2 + \sigma_1^2)}{\sigma_x^2 \sigma_1^2 + \sigma_x^2 \sigma_N^2 + \sigma_N^2 \sigma_1^2},$$

which is indeed the optimal value. It remains to show that the type-one error can be made arbitrarily small. This is straightforward by observing that each length-$n$ ECDQ quantization is independent of the others, and by the law of large numbers, when $m$ is sufficiently large, with high probability the sample average concentrates near its expected value, which is zero under hypothesis $H_0$. It is clear that choosing a sufficiently small but positive $\delta$ can drive the type-one error arbitrarily small when $m$ is large.

The above method can be used to bound the second stage error exponent $E_2$ by substituting quantization noise $\boldsymbol{N_2}$ similarly with an appropriate Gaussian random vector; the details are thus omitted. We note that strictly speaking, a system based on ECDQ is not a fixed-rate-coded deterministic system, thus it is not within the problem definition. Nevertheless, this randomized system can indeed be used to assert the existence of a fixed-rate and deterministic system of the same performance; see [29] for details.

## VI. Proof of the Converse for Theorem 5

In this section the converse proof of Theorem 5 is given by generalizing the image size characterization approach taken by Csiszár and Körner [12]. Since this proof relies heavily on the methods of types, the blowing-up lemma and some related concepts, we provide a brief review on these results in the Appendix. More details on the method of types can be found in [12]. In the remainder of this section we assume the readers' familiarity with Section 1.2, 1.5 and 2.1 of [12]; familiarity with Section 3.3 will also be helpful.

### A. Two lemmas

For a given probability distribution $P_{XY}$ which induces the channel $V^n : \mathcal{X}^n \to \mathcal{Y}^n$, the set $B \subseteq \mathcal{Y}^n$ is called an $\eta$-image of the set $A \subseteq \mathcal{X}^n$ over the channel $V^n$ if $V^n(B|x^n) \ge \eta$ for every $x^n \in A$ (see [12] page 101). The collection of $\eta$-images of the set $A$ is denoted as $\mathcal{B}(A, \eta)$. The following quantity is related to the minimum type-two error probability associated with set $A$

$$k_{V^n}(A, Q, \eta) = \frac{\min_{B \in \mathcal{B}(A,\eta)} Q_{XY}^n(A, B)}{P_X^n(A)}$$

where $Q_{XY}$ is the alternative hypothesis distribution; for the test against independence problem, since the alternative hypothesis is independence, we have

$$\frac{Q_{XY}^n(A, B)}{P_X^n(A)} = \frac{P_X^n(A) P_Y^n(B)}{P_X^n(A)} = P_Y^n(B).$$

In the sequel, only this case will be considered, and thus $k_{V^n}(A, Q, \eta)$ is simply written as $k_{V^n}(A, \eta)$. Note that $k_{V^n}(A, Q, \eta)$ is a generalization of the minimum cardinality of the $\eta$-images in [12], which was used to prove the channel coding theorem.

The following two lemmas are important for the converse proof. The first lemma essentially states that $\frac{1}{n}\log k_{V^n}(A,\eta)$ is independent of $\eta$ for sufficiently large $n$, while the second lemma provides a way to bound this quantity. Denote the letter $y \in \mathcal{Y}$ with the minimum probability in $P_y$ as $y_{\min}$, which is strictly positive as assumed, and define $\tau \triangleq -\log P_Y(y_{\min})$.

***Lemma 2:*** For every $\delta, \epsilon', \epsilon'' \in (0,1)$, we have for any set $A \subseteq \mathcal{X}^n$

$$\left|\frac{1}{n}\log k_{V^n}(A,\epsilon') - \frac{1}{n}\log k_{V^n}(A,\epsilon'')\right| < \delta,$$

whenever $n \geq n_0(|\mathcal{X}|, |\mathcal{Y}|, P_Y(y_{\min}), \delta, \epsilon', \epsilon'')$.

*Proof:* Suppose $\epsilon' > \epsilon''$. Clearly we have

$$\frac{1}{n}\log k_{V^n}(A,\epsilon') \geq \frac{1}{n}\log k_{V^n}(A,\epsilon'').$$

Let $B$ be an $\epsilon''$-image of $A$ which achieves $k_{V^n}(A,\epsilon'')$. Then by the blowing-up Lemma A-3, there exists a sequence $l_n$ with $\frac{l_n}{n} \to 0$ such that for sufficiently large $n > n_0(|\mathcal{X}|, |\mathcal{Y}|, \epsilon', \epsilon'')$

$$V^n(\Gamma^{l_n}B|x^n) \geq \epsilon' \quad \text{if} \quad V^n(B|x^n) \geq \epsilon'',$$

where $\Gamma^{l_n}B$ is the Hamming $l$-neighbourhood of $B$ (see (A-1)). This means $\Gamma^{l_n}B$ is an $\epsilon'$-image of $A$, and it implies that

$$k_{V^n}(A,\epsilon') \leq P_Y^n(\Gamma^{l_n}B).$$

Take this sequence of $\{l_n\}$ as that in Lemma A-2, then for sufficiently large $n > n_1(|\mathcal{X}|, |\mathcal{Y}|, P_Y(y_{\min}), \delta, \epsilon', \epsilon'')$, we have that

$$\frac{1}{n}\log P_Y^n(\Gamma^{l_n}B) - \frac{1}{n}\log P_Y^n(B) \leq \delta.$$

and it follows that

$$\begin{aligned}\frac{1}{n}\log k_{V^n}(A,\epsilon') &\leq \frac{1}{n}\log P_Y^n(\Gamma^{l_n}B)\\ &\leq \frac{1}{n}\log P_Y^n(B) + \delta = \frac{1}{n}\log k_{V^n}(A,\epsilon'') + \delta,\end{aligned}$$

which completes the proof. ■

***Lemma 3:*** For any set $A \subseteq \mathcal{X}^n$, consider a random vector $\hat{X}^n = (\hat{X}_1, \hat{X}_2, ..., \hat{X}_n)$ distributed over $A$ and let the random vector $\hat{Y}^n = (\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_n)$ be connected with $\hat{X}^n$ by the channel $V^n : \mathcal{X}^n \to \mathcal{Y}^n$, which is induced by $P_{XY}$. Then for every $\delta > 0$, $0 < \eta < 1$, we have

$$\frac{1}{n}D(P_{\hat{Y}^n}||P_Y^n) + \delta \geq -\frac{1}{n}\log k_{V^n}(A,\eta),$$

whenever $n \geq n_0(|\mathcal{X}|, |\mathcal{Y}|, P_Y(y_{\min}), \delta, \eta)$.

*Proof:* In light of Lemma 2, we only need to show that there exists an $\eta_0 = \eta_0(|\mathcal{Y}|, P_Y(y_{\min}), \delta)$ such that

$$\frac{1}{n}D(P_{\hat{Y}^n}||P_Y^n) + \delta \geq -\frac{1}{n}\log k_{V^n}(A,\eta_0),$$

if $n \geq n_1(|\mathcal{Y}|, P_Y(y_{\min}), \delta)$.

Let $B \in \mathcal{Y}^n$ be an $\eta_0$-image of $A$ that achieves $k_{V^n}(A,\eta_0)$. Then by the data-processing inequality for divergence, we have

$$D(P_{\hat{Y}^n}||P_Y^n) \geq \alpha\log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta},$$

where

$$\alpha \triangleq P_{\hat{Y}^n}(B), \quad \beta \triangleq P_Y^n(B) = k_{V^n}(A,\eta_0).$$

Since $B$ is an $\eta_0$-image of $A$, we have

$$\begin{aligned}\alpha &= \sum_{x^n \in A} P_{\hat{X}^n}(x^n)\Pr\{\hat{Y}^n \in B | \hat{X}^n = x^n\}\\ &\geq \sum_{x^n \in A} P_{\hat{X}^n}(x^n)\eta_0 = \eta_0.\end{aligned}$$

Thus we have

$$\frac{1}{n}D(P_{\hat{Y}^n}||P_Y^n) \geq -\frac{H_b(\alpha)}{n} - \frac{\eta_0}{n}\log k_{V^n}(A,\eta_0), \tag{12}$$

where again $H_b(\cdot)$ is the binary entropy function.

Notice the following simple fact

$$\begin{aligned}&D(P_{\hat{Y}^n}||P_Y^n)\\ &= \sum_{y^n \in \mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\log\frac{P_{\hat{Y}^n}(y^n)}{P_Y^n(y^n)}\\ &= -H(\hat{Y}^n) - \sum_{y^n \in \mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\log P_Y^n(y^n)\\ &\leq -\sum_{\hat{y}^n \in \mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\log P_Y^n(y_{\min}^n) = n\tau,\end{aligned}$$

where $\tau$ was defined before Lemma 2.

It follows from (12) that

$$\begin{aligned}&-\frac{1}{n}\log k_{V^n}(A,\eta_0)\\ &\leq \frac{1}{n}D(P_{\hat{Y}^n}||P_Y^n) + \frac{H_b(\alpha)}{n\eta_0} + [\frac{1}{n\eta_0} - \frac{1}{n}]D(P_{\hat{Y}^n}||P_Y^n)\\ &\leq \frac{1}{n}D(P_{\hat{Y}^n}||P_{Y^n}) + \frac{1}{n\eta_0} + [\frac{1}{\eta_0} - 1]\tau.\end{aligned}$$

By choosing an appropriate $\eta_0$, e.g. $\eta_0 = \frac{2\tau+\delta}{2(\delta+\tau)}$, the following inequality is satisfied

$$\frac{1}{n\eta_0} + [\frac{1}{\eta_0} - 1]\tau \leq \delta$$

whenever $n \geq n_1(|\mathcal{Y}|, P_Y(y_{\min}), \delta)$ and the proof is complete. ■

### *B. Converse proof of Theorem 5*

Now we are ready to prove the converse of Theorem 5, which establishes the complete characterization of $\mathcal{R}_{ht}(\epsilon_1, \epsilon_2)$. We shall be considering several probability distributions in this proof, and the region $\mathcal{R}_{ht}^*$ will be written as $\mathcal{R}_{ht}^*(X,Y)$, in order to emphasize the dependence on the particular distribution in consideration. Only the case $\epsilon_1 + \epsilon_2 < 1$ needs to be considered, since for the other case the strong converse result apparently follows from that in [9]. Let the channel $V^n : X^n \to Y^n$ be that induced by $P_{XY}$. We will take the Delta-convention in [12] (p. 34) and suppress the dependence of all the small quantities on $n$. Note also that the sets defined below such as $A_1, A_2, B_1, B_2, C$ are all in fact sequences of sets indexed by $n$, however we are suppressing it for simplicity. In this subsection alone, we use $t$ for the time index, and $i$ and $j$ for the encoding function values.

*Proof of the converse for Theorem 5:*

For any two encoding functions $f_1, f_2 \in \mathcal{F}_n$ with two sets $A_1 \subseteq f_1(\mathcal{X}^n) \times \mathcal{Y}_n$ and $A_2 \subseteq f_1(\mathcal{X}^n) \times f_2(\mathcal{X}^n) \times \mathcal{Y}_n$ such that

$$P_{f_1(X^n)Y^n}(A_1) \geq 1-\epsilon_1,\ P_{f_1(X^n)f_2(X^n)Y^n}(A_2) \geq 1-\epsilon_2,$$

we may assume that

$$A_1 = \bigcup_{i=1}^{|f_1|} i \times G_i, \quad G_i \subseteq \mathcal{Y}^n, \quad i = 1,2,...,|f_1|,$$
$$A_2 = \bigcup_{i=1,j=1}^{|f_1|,|f_2|} (i,j) \times G_{i,j},$$
$$G_{i,j} \subseteq \mathcal{Y}^n, \quad i = 1,2,...,|f_1|, \quad j = 1,2,...,|f_2|.$$

Define the following sets

$$B_1 = \left\{ x^n : x^n \in \mathcal{X}^n, V^n(G_{f_1(x^n)}|x^n) \geq \frac{1-\epsilon_1-\epsilon_2}{1+3\epsilon_1-\epsilon_2} \right\}$$
$$B_2 = \Bigg\{ x^n : x^n \in \mathcal{X}^n, V^n(G_{f_1(x^n),f_2(x^n)}|x^n) \geq \frac{1-\epsilon_1-\epsilon_2}{1+3\epsilon_2-\epsilon_1} \Bigg\}.$$

Since we have

$$\begin{aligned}
1-\epsilon_1 &\leq P_{f_1(X^n)Y^n}(A_1) \\
&= \sum_{x^n \in \mathcal{X}^n} P_X^n(x^n) V^n(G_{f_1(x^n)}|x^n) \\
&= \sum_{x^n \in B_1} P_X^n(x^n) V^n(G_{f_1(x^n)}|x^n) \\
&\qquad + \sum_{x^n \in B_1^c} P_X^n(x^n) V^n(G_{f_1(x^n)}|x^n) \\
&\leq P_X^n(B_1) + (1-P_X^n(B_1)) \frac{1-\epsilon_1-\epsilon_2}{1+3\epsilon_1-\epsilon_2},
\end{aligned}$$

it follows that

$$P_X^n(B_1) \geq \frac{3-3\epsilon_1+\epsilon_2}{4}.$$

Similarly, we have that

$$P_X^n(B_2) \geq \frac{3-3\epsilon_2+\epsilon_1}{4}.$$

This implies that

$$\begin{aligned}
P_X^n(B_1 \cap B_2) &\geq \frac{3-3\epsilon_1+\epsilon_2}{4} + \frac{3-3\epsilon_2+\epsilon_1}{4} - 1 \\
&= \frac{1-\epsilon_1-\epsilon_2}{2} > 0.
\end{aligned}$$

By the property of typical sequences given in Lemma A-1, it follows that for any $\delta''$ such that

$$0 < \delta'' < \frac{1-\epsilon_1-\epsilon_2}{2},$$

we have for any $\delta' > 0$

$$P_X^n(B_1 \cap B_2 \cap T_{[X]\delta'}^n) \geq \delta'',$$

whenever $n \geq n_0(|\mathcal{X}|, \delta'')$. Next we find a single type in the intersection $B_1 \cap B_2 \cap T_{[X]\delta'}^n$ with the maximum probability, and denote this type as $P_0$. Since there are less than $(n+1)^{|\mathcal{X}|}$ types in total, it follows that

$$P_X^n(B_1 \cap B_2 \cap T_{P_0}^n) \geq \frac{\delta''}{(n+1)^{|\mathcal{X}|}}. \tag{13}$$

From this point on we essentially consider only this single type. For simplicity, define $C \triangleq B_1 \cap B_2 \cap T_{P_0[X]}^n$. Furthermore, for any $x^n \in T_{[X]\delta'}^n$, we have that

$$P_X^n(x^n) \leq \exp(-n(H(X)-\delta_1)), \tag{14}$$

where $\delta_1 \to 0$ as $\delta' \to 0$, it follows

$$\frac{1}{n}\log|C| \geq H(X) - \delta_2, \tag{15}$$

where $\delta_2 \to 0$ as $\delta' \to 0$.

The functions $f_1$ and $f_2$ clearly partition the set $C$ into $|f_1||f_2|$ non-intersecting subsets; denote those sets as $C_{i,j}$. Assign a uniform distribution $P_{\hat{X}^n}$ onto the set $C$, and denote the resulting random variable $f_1(\hat{X}^n)$ as $T_1$, where $\hat{X}^n$ is the random variable uniformly distributed on $C$ by distribution $P_{\hat{X}^n}$; similarly, denote $f_2(\hat{X}^n)$ as $T_2$. Let $\hat{Y}^n$ be connected with $\hat{X}^n$ by the channel $V^n : \mathcal{X}^n \to \mathcal{Y}^n$. Apparently $(T_1, T_2) \leftrightarrow \hat{X}^n \leftrightarrow \hat{Y}^n$ forms a Markov chain.

It is clear that we have

$$\begin{aligned}
\log|f_1| &\geq H(T_1) = I(T_1; \hat{X}^n) \\
&= H(\hat{X}^n) - H(\hat{X}^n|T_1) \\
&= \log|C| - H(\hat{X}^n|T_1) \\
&\overset{(a)}{\geq} n(H(X)-\delta_2) - H(\hat{X}^n|T_1) \\
&= nH(X) - n\delta_2 - \sum_{t=1}^n H(\hat{X}_t|T_1\hat{X}_t^-),
\end{aligned} \tag{16}$$

where we have used (15) in (a). Similarly, we have

$$\log|f_1| + \log|f_2| \geq nH(X) - n\delta_2 - \sum_{t=1}^n H(\hat{X}_t|T_1T_2\hat{X}_t^-). \tag{17}$$

Notice that $G_i$ is in fact a $\frac{1-\epsilon_1-\epsilon_2}{1+3\epsilon_1-\epsilon_2}$-image for the set $C_i \triangleq \bigcup_j C_{i,j}$. We can now bound the type-two error at the first stage as follows

$$\begin{aligned}
\beta_1 &\geq \sum_{x^n \in C} P_X^n(x^n) P_Y^n(G_{f_1(x^n)}) \\
&= \sum_{i=1}^{|f_1|} P_X^n(C_i) P_Y^n(G_i) \\
&\geq \sum_{i=1}^{|f_1|} P_X^n(C_i) k_{V^n}\left(C_i, \frac{1-\epsilon_1-\epsilon_2}{1+3\epsilon_1-\epsilon_2}\right) \\
&\geq \sum_{i=1}^{|f_1|} P_X^n(C_i) \exp\left(-D(\hat{Y}^n||Y^n|T_1=i) - n\delta\right),
\end{aligned}$$

where the last step we used Lemma 3; note that conditioning is needed here for the divergence term, however it is related only to the $\hat{Y}^n$ term by limiting $T_1 = f_1(\hat{X}^n) = i$. It further

follows

$$\begin{aligned}
\beta_1 &\geq \sum_{i=1}^{|f_1|} P_X^n(C_i) \exp\Big(-D(\hat{Y}^n||Y^n|T_1=i) - n\delta\Big) \\
&\overset{(a)}{=} P_X^n(C) \sum_{i=1}^{|f_1|} \frac{|C_i|}{|C|} \exp\Big(-D(\hat{Y}^n||Y^n|T_1=i) - n\delta\Big) \\
&\overset{(b)}{=} P_X^n(C) \sum_{i=1}^{|f_1|} P_{\hat{X}^n}(C_i) \exp\Big(-D(\hat{Y}^n||Y^n|T_1=i) - n\delta\Big) \\
&\overset{(c)}{\geq} \exp(-n\delta) \\
&\qquad \times P_X^n(C) \exp\left(-\sum_{i=1}^{|f_1|} P_{\hat{X}^n}(C_i) D(\hat{Y}^n||Y^n|T_1=i)\right),
\end{aligned}$$

where (a) and (b) are due to the fact that the set $C$ consists of sequences of the same type and $P_{\hat{X}^n}$ is a uniform distribution on $C$, and in (c) we used the convexity of function $\exp(\cdot)$. It is worth noting that the bounding above turns out to be tight suggests that the distribution of $\hat{Y}^n$ given $T_1 = i$ is approximately the same for each value of $i$; this in turn implies that the set $C$ is partitioned in an approximately uniform fashion into sets of similar structure by $f_1$ (and $f_2$). Now it follows

$$-\frac{1}{n}\log\beta_1 \leq \frac{1}{n}\sum_{i=1}^{|f_1|} P_{\hat{X}^n}(C_i) D(\hat{Y}^n||Y^n|T_1=i) + \delta_3$$

where $\delta_3 = \delta - \frac{1}{n}\log\frac{\delta''}{(n+1)^{|\mathcal{X}|}}$, and we used the fact $P_X^n(C) \geq P_C^n(B_1 \cap B_2 \cap C)$ and (13). We continue the chain of inequalities as follows

$$\begin{aligned}
&-\frac{1}{n}\log\beta_1 - \delta_3 \leq \frac{1}{n}\sum_{i=1}^{|f_1|} P_{\hat{X}^n}(C_i) D(\hat{Y}^n||Y^n|T_1=i) \\
&= \frac{1}{n}\sum_{i=1}^{|f_1|} P_{\hat{X}^n}(C_i) \sum_{y^n\in\mathcal{Y}^n} P_{\hat{Y}^n|T_1=i}(y^n) \log\frac{P_{\hat{Y}^n|T_1=i}(y^n)}{P_Y^n(y^n)} \\
&= \frac{1}{n}\sum_{i=1}^{|f_1|} \sum_{y^n\in\mathcal{Y}^n} P_{\hat{X}^n\hat{Y}^n}(C_i, y^n) \log P_{\hat{Y}^n|T_i=i}(y^n) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{|f_1|} \sum_{y^n\in\mathcal{Y}^n} P_{\hat{X}^n\hat{Y}^n}(C_i, y^n) \log P_Y^n(y^n) \\
&= -\frac{1}{n}H(\hat{Y}^n|T_1) - \frac{1}{n}\sum_{y^n\in\mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\log P_Y^n(y^n) \qquad (18)
\end{aligned}$$

Because $P_Y^n$ is a product distribution, we have

$$\begin{aligned}
&\sum_{y^n\in\mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\log P_Y^n(y^n) \\
&= \sum_{y^n\in\mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\sum_{t=1}^{n}\log P_Y(y_t) \\
&= \sum_{y^n\in\mathcal{Y}^n}\sum_{t=1}^{n} P_{\hat{Y}^n}(y^n)\log P_Y(y_t) \\
&= \sum_{t=1}^{n}\sum_{y^n\in\mathcal{Y}^n} P_{\hat{Y}^n}(y^n)\log P_Y(y_t) \\
&= \sum_{t=1}^{n}\sum_{y_t\in\mathcal{Y}}\left(\sum_{y^n:y^n(t)=y_t} P_{\hat{Y}^n}(y^n)\right)\log P_Y(y_t) \\
&= \sum_{t=1}^{n}\sum_{y_t\in\mathcal{Y}} P_{\hat{Y}_t}(y_t)\log P_Y(y_t) \\
&= \sum_{t=1}^{n}[-H(\hat{Y}_t) - D(\hat{Y}_t||Y_t)].
\end{aligned}$$

Resuming from (18) it follows that

$$\begin{aligned}
&-\frac{1}{n}\log\beta_1 - \delta_3 \\
&\leq -\frac{1}{n}H(\hat{Y}^n|T_1) + \frac{1}{n}\sum_{t=1}^{n}[H(\hat{Y}_t) + D(\hat{Y}_t||Y_t)] \\
&= \frac{1}{n}\sum_{t=1}^{n}[H(\hat{Y}_t) - H(\hat{Y}_t|T_1\hat{Y}_t^-) + D(\hat{Y}_t||Y_t)] \\
&\leq \frac{1}{n}\sum_{t=1}^{n}[H(\hat{Y}_t) - H(\hat{Y}_t|T_1\hat{X}_t^-\hat{Y}_t^-) + D(\hat{Y}_t||Y_t)] \\
&\overset{(a)}{\leq} \frac{1}{n}\sum_{t=1}^{n}[H(\hat{Y}_t) - H(\hat{Y}_t|T_1\hat{X}_t^-) + D(\hat{Y}_t||Y_t)] \\
&= \frac{1}{n}\sum_{t=1}^{n}[I(\hat{Y}_t; T_1\hat{X}_t^-) + D(\hat{Y}_t||Y_t)],
\end{aligned}$$

where (a) is due to the Markov string $\hat{Y}_t \leftrightarrow (T_1\hat{X}_t^-) \leftrightarrow \hat{Y}_t^-$. By a similar manner, we can get

$$-\frac{1}{n}\log\beta_2 \leq \frac{1}{n}\sum_{t=1}^{n}[I(\hat{Y}_t; T_1T_2\hat{X}_t^-) + D(\hat{Y}_t||Y_t)] + \delta_3. \quad (19)$$

Now introduce a random variable $J$ uniformly distributed over the set $I_n$, and independent of $T_1, T_2, \hat{X}^n, \hat{Y}^n$. For convenience introduce the following notations,

$$\overline{X} = \hat{X}_J, \quad \overline{Y} = \hat{Y}_J, \quad U = (J, T_1, \hat{X}_J^-), \quad V = (J, T_2),$$

and it follows from (16) and (17) that

$$\begin{aligned}
\frac{1}{n}\log|f_1| &\geq H(X) - \delta_2 - H(\overline{X}|U) \\
\frac{1}{n}\log|f_1| + \frac{1}{n}\log|f_2| &\geq H(X) - \delta_2 - H(\overline{X}|UV)
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}[I(\hat{Y}_t;T_1\hat{X}_t^-)+D(\hat{Y}_t||Y_t)]\\
&=I(\hat{Y}_J;U|J)+\frac{1}{n}\sum_{t=1}^{n}\sum_{y\in\mathcal{Y}}P_{\hat{Y}_J|J=t}(y)\log\frac{P_{\hat{Y}_J|J=t}(y)}{P_{Y_J|J=t}(y)}\\
&\overset{(a)}{=} I(\hat{Y}_J;U|J)\\
&\qquad+\frac{1}{n}\sum_{t=1}^{n}\sum_{y\in\mathcal{Y}}P_{\hat{Y}_J|J=t}(y)\log\frac{P_{\hat{Y}_J|J=t}(y)P_{\hat{Y}_J}(y)}{P_{Y_J}(y)P_{\hat{Y}_J}(y)}\\
&=I(\hat{Y}_J;U|J)-H(\hat{Y}_J|J)+D(\hat{Y}_J||Y_J)+H(\hat{Y}_J)\\
&=I(\hat{Y}_J;U|J)+I(\hat{Y}_J;J)+D(\hat{Y}_J||Y_J)\\
&=I(\hat{Y}_J;U,J)+D(\hat{Y}_J||Y_J)\\
&\overset{(b)}{=} I(\hat{Y}_J;U)+D(\hat{Y}_J||Y_J)\\
&=I(\overline{Y};U)+D(\overline{Y}||Y),
\end{aligned}$$

where (a) is because $P_{Y_J}$ is in fact independent of $J$, and (b) is because $U=(J,T_1,\hat{X}_J^-)$. Similarlly

$$\frac{1}{n}\sum_{t=1}^{n}[I(\hat{Y}_t;T_1T_2\hat{X}_t^-)+D(\hat{Y}_t||Y_t)]=I(\overline{Y};UV)+D(\overline{Y}||Y).$$

And it follows that

$$\begin{aligned}
-\frac{1}{n}\log\beta_1 &\leq I(\overline{Y};U)+D(\overline{Y}||Y)+\delta_3\\
-\frac{1}{n}\log\beta_2 &\leq I(\overline{Y};UV)+D(\overline{Y}||Y)+\delta_3.
\end{aligned}$$

Clearly,

$$P_{\overline{X}}(x)=\frac{1}{n}\sum_{t=1}^{n}P_{\hat{X}_t}(x)=\frac{1}{n|C|}\sum_{\boldsymbol{x}\in C}N(x|\boldsymbol{x}),\quad x\in\mathcal{X} \tag{20}$$

and by the definitions we have

$$P_{\overline{Y}|\overline{X}}=P_{Y|X}.$$

Furthermore, it is straightforward to check the Markov string $\overline{Y}\leftrightarrow\overline{X}\leftrightarrow(UV)$.

So far we have proved the following

$$\begin{aligned}
&\mathcal{R}_{ht}(\epsilon_1,\epsilon_2)\in\mathcal{R}_{ht}^*(\overline{X},\overline{Y})\\
&\quad+[H(X)-H(\overline{X}),H(X)-H(\overline{X}),D(\overline{Y}||Y),D(\overline{Y}||Y)],
\end{aligned} \tag{21}$$

for any $\epsilon_1,\epsilon_2\in(0,1)$ and $\epsilon_1+\epsilon_2<1$. The proof can be completed by a continuity argument, if $P_{\overline{XY}}$ is sufficiently close to $P_{XY}$.

By (20), and the fact that $C\subseteq T^n_{[X]\delta'}$, we have

$$|P_X(x)-P_{\overline{X}}(x)|<\delta',\quad x\in\mathcal{X}.$$

as well as $P_{\overline{Y}|\overline{X}}=P_{Y|X}$. By the uniform continuity of involved information quantities, it follows that if $\delta'$ is sufficiently small, for every point $(\overline{R_1},\overline{R_2},\overline{E_1},\overline{E_2})\in\mathcal{R}_{ht}^*(\overline{X},\overline{Y})$, there exists a point $(R_1,R_2,E_1,E_2)\in\mathcal{R}_{ht}^*(X,Y)$ that is arbitrarily close to it (see [12] p. 322 for details). Note that here we need to use again the positivity of $P_Y$ for the continuity to hold for $D(\overline{Y}||Y)$. The proof is complete by asserting that such $\delta'$ indeed exists for any sufficiently large $n$. ■

## VII. CONCLUSION

We investigated two closely related problems, namely successive refinement hypothesis testing and successive refinement lossless one-helper problem. It was shown that the the rate-exponent region of the former and rate regions of the latter are congruent to each other. The unified approach facilitates the treatment and provides several non-trivial results. We focus on the SR-HT problem, and a strong converse result is proved for this problem. Gaussian problem was investigated in some depth for the SR-HT problem. Moreover, a new operational meaning of the information bottleneck method was revealed by connection to the problems being considered, which is more intuitive than previous given in the literature.

We believe the entropy characterization problem extracted from these problems is fundamentally important, which has not been fully explored. Future research along this direction may provide results in other multi-terminal information theoretical problems.

## APPENDIX

***Definition** A-1:* Given a set $B\in\mathcal{Y}^n$, the Hamming $l$-neighborhood of $B$ is defined as the set

$$\Gamma^lB\triangleq\{\boldsymbol{y}:\boldsymbol{y}\in\mathcal{Y}^n,d_H(\{\boldsymbol{y}\},B)\leq l\}, \tag{A-1}$$

where $d_H(B,C)$ denotes the Hamming metric between two sets $B$ and $C$ by extending the usual Hamming distance of two sequences $d_h(\cdot,\cdot)$ as

$$d_H(B,C)\triangleq\min_{\boldsymbol{y}\in B,\hat{\boldsymbol{y}}\in C}d_h(\boldsymbol{y},\hat{\boldsymbol{y}})$$

***Lemma** A-1:* For any $\delta>0$, there exists a sequence $\epsilon_n\to0$ depending only on $|\mathcal{X}|$ so that for every distribution $P$ on $\mathcal{X}$

$$P^n(T^n_{[P]\delta})\geq1-\epsilon_n.$$

***Lemma** A-2:* Given a sequence of positive integers $\{l_n\}$ with $\frac{l_n}{n}\to0$ and a distribution $P$ on $\mathcal{Y}$ with positive probabilities, there exists a sequence $\epsilon_n\to0$ depending only on $\{l_n\}$, $|\mathcal{Y}|$ and $\min_{y\in\mathcal{Y}}P(y)$ such that for every $B\subseteq\mathcal{Y}^n$

$$\begin{aligned}
0\leq\frac{1}{n}\log|\Gamma^{l_n}B|-\frac{1}{n}\log|B|\leq\epsilon_n,\\
0\leq\frac{1}{n}\log P^n(\Gamma^{l_n}B)-\frac{1}{n}\log P^n(B)\leq\epsilon_n.
\end{aligned}$$

***Lemma** A-3 (Blowing up):* To any finite sets $\mathcal{X}$ and $\mathcal{Y}$ and sequence $\epsilon_n\to0$, there exists a sequence of positive integers $l_n$ with $\frac{l_n}{n}\to0$ and a sequence $\eta_n\to1$ such that for every stochastic matrix $W:\mathcal{X}\to\mathcal{Y}$ and every $n$, $\boldsymbol{x}\in\mathcal{X}^n$, $B\subseteq\mathcal{Y}^n$

$$W^n(B|\boldsymbol{x})\geq\exp(-n\epsilon_n)\text{ implies }W^n(\Gamma^{l_n}B|\boldsymbol{x})\geq\eta_n.$$

## REFERENCES


[1] V. N. Koshelev, "Hierarchical coding of discrete sources," *Probl. Pered. Inform.*, vol. 16, no. 3, pp. 31-49, 1980.

[2] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Information Theory*, vol. 37, no. 2, pp. 269-275, Mar. 1991.

[3] B. Rimoldi, "Successive refinement of information: Characterization of achievable rates," *IEEE Trans. Information Theory*, vol. 40, no. 1, pp. 253-259, Jan. 1994.

[4] M. Effros, "Distortion-rate bounds for fixed- and variable-rate multiresolution source codes," *IEEE Trans. Information Theory*, vol. 45, no. 6, pp. 1887-1910, Sep. 1999.

[5] M. Effros, "Universal multiresolution source codes," *IEEE Trans. Information Theory*, vol. 47, no. 6, pp. 2113-2129, Sep. 2001.

[6] E. Tuncel and K. Rose, "Error exponents in scalable source coding," *IEEE Trans. Information Theory*, vol. 49, no. 1, pp. 289-296, Jan. 2003.

[7] E. Tuncel and K. Rose, "Computation and analysis of the $n$-layer scalable rate-distortion function," *IEEE Trans. Information Theory*, vol. 49, no. 5, pp. 1218-1230, May. 2003.

[8] E. Tuncel and K. Rose, "Additive successive refinement," *IEEE Trans. Information Theory*, vol. 49, no. 8, pp. 1983-1991, Aug. 2003.

[9] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Information Theory*, vol. 32, no. 4, pp. 533-542, Jul. 1986.

[10] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2300-2324, Oct. 1998.

[11] R. Ahlswede, "Coloring hypergraphs: A new approach to multi-user source coding: Part II," *J. Combinatorics, Infom. Syst. Sci.*, vol. 5, pp. 220-268, 1980.

[12] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press, New York, 1981.

[13] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Information Theory*, vol. 21, pp. 294-300, May 1975.

[14] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Information Theory*, vol. 21, no. 6, pp. 629-637, Nov. 1975.

[15] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. Information Theory*, vol. 43, no. 6, pp. 1912-1923, Nov. 1997.

[16] T. Berger and R.W. Yeung, "Multiterminal source encoding with encoder breakdown," *IEEE Trans. Information Theory*, vol. 35, no. 2, pp. 237-244, Mar. 1989.

[17] J. Chen and T. Berger, "Robust distributed source coding," *submitted for publication*.

[18] A. D. Wyner, "A theorem on the entropy of certain binary sequences and application: II," *IEEE Trans. Information Theory*, vol. 19, pp. 772–777, Nov. 1973.

[19] E. Tuncel, "Capacity/storage tradeoff in high-dimensional identification systems," in *Proc. IEEE International symposium on information theory*, Seattle, WA, USA, pp. 1929-1933, Jul. 2006.

[20] E. Tuncel, "Capacity/storage tradeoff in high-dimensional identification systems," *submitted to publication*.

[21] M. B. Westover and J. A. O'Sullivan, "Towards an information theoretic framework for object recognition," in *Proc. IEEE International symposium on information theory*, Chicago, Illinois, USA, p.219, Jun.-Jul 2004.

[22] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Information Theory*, vol. 54, no. 1, pp. 299-320, Jan. 2008.

[23] J. A. O'Sullivan, N. Singla, and M. B. Westover, "Successive refinement for pattern recognition," in *Proc. IEEE Information Theory Workshop*, Punta del Este, Uruguay, Mar. 2006, pp. 141-145.

[24] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, Allerton House, University of Illinois, USA, Sep. 1999.

[25] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proceedings of the 16'th Conference on Computational Theory (COLT)*, 2003, pp. 595-609.

[26] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proceedings of 2007 IEEE International Symposium on Information Theory*, Nice, France, Jun. 2007.

[27] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *Proceeding SIGIR'02, 25th ACM International Conference on Research and Development of Information Retireval.*, Tampere, pp. 129-136, Finland, 2002.

[28] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493-507, 1952.

[29] C. Tian and J. Chen, "Successive refinement for hypothesis testing, one-helper problem and pattern recognition," *McMaster University Technical report*, 2008.

[30] T. M. Cover and J. A. Thomas, *Elements of information theory*, New York: Wiley, 1991.

[31] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information source," *IEEE Trans. Information Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[32] A. Sgarro, "Source coding with side information at several decoders," *IEEE Trans. Information Theory*, vol. 23, no. 2, pp. 179–182, Mar. 1977.

[33] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Information Theory*, vol. 31, no. 6, pp. 727–734, Nov. 1985.

[34] C. Tian and S. Diggavi, "Side information scalable source coding," *submitted for publication*.

[35] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. Journal*, vol. 59, pp. 1909–1921, Dec. 1980.

[36] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Information Theory*, vol. 41, no. 1, pp. 141–154, Jan. 1995.

[37] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Information Theory*, vol. 42, no. 4, pp. 1152–1159, Jul. 1996.

[38] R. Zamir and M. Feder, "Information rates of pre/post filtered dithered quantizers," *IEEE Trans. Information Theory*, vol. 42, no. 5, pp. 1340–1353, Sep. 1996.

[39] U. Erez and R. Zamir, "Achieving $\frac{1}{2}\log(1+\text{SNR})$ on the AWGN channel with lattice encoding and decoding," *IEEE Trans. Information Theory*, vol. 50, no. 10, pp. 2293–2314, Oct. 2004.

[40] T. Liu, P. Moulin, and R. Koetter, "On error exponents of modulo lattice additive noise channels," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 454–471, Feb. 2006.

[41] C. A. Rogers, *Packing and covering*, Cambridge, U.K.: Camdridge Univ. Press, 1964.

[42] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*, New York: Springer-Verlag, 1998.



**Chao Tian** (S'00, M'05) received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2000 and the M.S. and Ph. D. degrees in Electrical and Computer Engineering from Cornell University, Ithaca, NY in 2003 and 2005, respectively.

Dr. Tian was a postdoctoral researcher at Ecole Polytechnique Federale de Lausanne (EPFL) from 2005 to 2007. He joined AT&T Labs–Research, Florham Park, New Jersey in 2007, where he is now a Senior Member of Technical Staff. His research interests include multi-user information theory, joint source-channel coding, quantization design and analysis, as well as image/video coding and processing.

**Jun Chen** (S'03, M'06) received the B.S. degree with honors in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from Cornell University, Ithaca, NY in 2003 and 2005, respectively. He was a Postdoctoral Research Associate in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign, Urbana, IL from 2005 to 2006, and a Josef Raviv Memorial Postdoctoral Fellow at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY from 2006 to 2007. He is currently an Assistant Professor of Electrical and Computer Engineering at McMaster University, Hamilton, Ontario, Canada. He holds the Barber-Gennum Chair in Information Technology.