On the Redundancy of Slepian–Wolf Coding

Da-ke He, *Member, IEEE*, Luis A. Lastras-Montaño, *Senior Member, IEEE*, En-hui Yang, *Fellow, IEEE*, Ashish Jagmohan, and Jun Chen, *Member, IEEE*

***Abstract*—In this paper, the redundancy of both variable and fixed rate Slepian–Wolf coding is considered. Given any jointly memoryless source-side information pair $\{(X_i, Y_i)\}_{i=1}^{\infty}$ with finite alphabet, the redundancy $R^n(\epsilon_n)$ of variable rate Slepian–Wolf coding of X_1^n with decoder only side information Y_1^n depends on both the block length n and the decoding block error probability ϵ_n, and is defined as the difference between the minimum average compression rate of order n variable rate Slepian–Wolf codes having the decoding block error probability less than or equal to ϵ_n, and the conditional entropy $H(X|Y)$, where $H(X|Y)$ is the conditional entropy rate of the source given the side information. The redundancy of fixed rate Slepian–Wolf coding of X_1^n with decoder only side information Y_1^n is defined similarly and denoted by $R_F^n(\epsilon_n)$. It is proved that under mild assumptions about ϵ_n, $R^n(\epsilon_n) = d_v\sqrt{-\log\epsilon_n/n} + o(\sqrt{-\log\epsilon_n/n})$ and $R_F^n(\epsilon_n) = d_f\sqrt{-\log\epsilon_n/n} + o(\sqrt{-\log\epsilon_n/n})$, where d_f and d_v are two constants completely determined by the joint distribution of the source-side information pair. Since d_v is generally smaller than d_f, our results show that variable rate Slepian–Wolf coding is indeed more efficient than fixed rate Slepian–Wolf coding.**

***Index Terms*—Compression rate, error probability, lossless data compression, redundancy, side information, Slepian–Wolf coding, source coding.**

I. Introduction

Let (X, Y) be a pair of random variables taking values in finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ denote a sequence of independent copies of (X, Y). For brevity, the memoryless sources $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are sometimes referred to as the sources X and Y, respectively. Suppose that the source X is to be compressed without essential loss of information, and Y is available only to the decoder as a helper, or in other words, the side information. This problem of source coding with decoder only side information was first introduced and studied by Slepian and Wolf in their ground breaking paper [13], and is now commonly referred to as the Slepian–Wolf (SW) coding problem (with one encoder). It was shown in [13] that for any memoryless pair (X, Y), there exist SW codes that can compress X with the decoder-only side information Y arbitrarily close to $H(X|Y)$ (asymptotically) while still recovering $\{X_i\}_{i=1}^{\infty}$ with vanishing error probability ϵ, which is defined as the probability that the source's message differs from the reconstructed message at the decoder. In other words, one can do, asymptotically, as well as in the case where the side information Y is also available at the encoder. This result is one of the first and remains as one of the most influential results in network information theory ever obtained.

Beyond its significant impact on network information theory, SW coding has also attracted a lot of attention from a practical side; recently there have been significant advances in constructing practical SW codes [11], [7], [12], [14], [16], and researching their use in various applications (see [8] and the references therein). A natural consequence of this surge of interest in the practice of SW coding is a renewed interest in understanding better the fundamental limitations of finite block length codes from an information theoretical point of view. For this purpose, we define the redundancy $R_C^n(\epsilon_n)$ of an (order n) SW code C as $r(C) - H(X|Y)$, where ϵ_n denotes the decoding error probability of C, and $r(C)$ denotes the average compression rate in bits per letter resulting from using the code C to compress $X_1X_2\cdots X_n$. Further define

$$R^n(\epsilon_n) \triangleq \min_C R_C^n(\epsilon_n) \tag{1.1}$$

where the minimization is taken over all (order n) variable rate SW codes with decoding error probability no greater than ϵ_n. The quantity $R^n(\epsilon_n)$ is called the redundancy of variable rate SW coding of X_1^n with the decoder only side information Y_1^n and decoding block error probability ϵ_n. If the minimization in (1.1) is limited to the set of all (order n) fixed rate SW codes with decoding error probability no greater than ϵ_n, the resulting quantity (denoted by $R_F^n(\epsilon_n)$) is called the redundancy of fixed rate SW coding of X_1^n with the decoder only side information Y_1^n and decoding block error probability ϵ_n.

Previous research efforts in understanding the performance of finite block length SW codes have been mainly focused on fixed rate codes. Specifically, Wolfowitz's treatment [15] of SW coding demonstrates the existence of fixed rate codes that operate within $K(\epsilon_n)/\sqrt{n}$ of the conditional entropy for some function $K(\epsilon_n)$. Classic results by Gallager [6], Csiszár and Körner [5] (obtained also with K. Marton), describe good upper

Manuscript received February 14, 2008; revised June 01, 2009. Current version published November 20, 2009. The work of E.-h. Yang is supported in part by the Natural Sciences and Engineering Research Council of Canada under Grants RGPIN203035-02 and RGPIN203035-06, and by the Canada Research Chairs Program. The material in this paper was presented in part at the 2006 IEEE International Symposium on Information Theory (ISIT), Seattle, WA, the 2006 IEEE Information Theory Workshop, Chengdu, China, and the 2007 IEEE International Symposium on Information Theory (ISIT), Nice, France.

D.-k. He was with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA. He is now with RIM/SlipStream, Waterloo, ON N2L 5Z5, Canada (e-mail: dhe@rim.com).

L. A. Lastras-Montaño and A. Jagmohan are with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: lastrasl@us.ibm.com; ashishja@us.ibm.com).

E.-h. Yang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: ehyang@uwaterloo.ca).

J. Chen was with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA. He is now with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: junchen@ece.mcmaster.ca).

Communicated by E. Ordentlich, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2009.2032803

and lower bounds on the error exponent of fixed rate SW coding[1] (the best exponential rate of decay of the decoding error probability as a function of the rate provided). More recently, Baron *et al.* [2] studied $R_F^n(\epsilon_n)$ for a pair of uniform binary random variables (X,Y) connected through a binary symmetric channel.

In contrast, other than the fact that

$$R^n(\epsilon_n) \le R_F^n(\epsilon_n) \tag{1.2}$$

little progress has been made towards fully understanding the redundancy $R^n(\epsilon_n)$ of variable rate SW. The purpose of this paper is to characterize this quantity and $R_F^n(\epsilon_n)$, and hence gain insights into the performance of SW codes in practice. We shall show that under mild assumptions[2] about ϵ_n

$$R^n(\epsilon_n) = d_v\sqrt{-\log\epsilon_n/n} + o(\sqrt{-\log\epsilon_n/n}) \tag{1.3}$$
$$R_F^n(\epsilon_n) = d_f\sqrt{-\log\epsilon_n/n} + o(\sqrt{-\log\epsilon_n/n}) \tag{1.4}$$

where d_v and d_f are two constants completely determined by the joint distribution of (X,Y). The exact formulae to compute d_f and d_v are provided later in Sections II and III, respectively.

A couple of implications can be drawn immediately from (1.3) and (1.4). First, the redundancy of SW coding is significantly larger than that of conditional coding given the side information available to both the encoder and decoder, which is $O(1/n)$. This difference might explain why designing efficient SW codes, in comparison to designing conditional source codes, is so challenging. Second, the design of practical SW codes with finite block length is not simply a matter of approaching the conditional entropy rate $H(X|Y)$; instead, it is more about the tradeoff among the compression rate, decoding error probability, and block length. Third, since d_v is strictly less than d_f in general, variable rate SW coding is indeed more efficient than fixed rate SW coding for finite block length, which could not be revealed by the first order performance analysis.

The rest of the paper is organized as follows. In Section II, we introduce the concept of intrinsic entropy and prove some preliminary results which will facilitate our later discussions. Section III is devoted to establishing lower bounds to $R^n(\epsilon_n)$ and $R_F^n(\epsilon_n)$. In Section IV, we show that the lower bounds in Section III are indeed tight by establishing matching upper bounds. In Section V, we compare the compression performance of variable rate SW coding with that of fixed rate Sleapian–Wolf coding for finite block lengths. Final discussions are given in Section VIII.

[1]The tradeoff between compression rate and decoding error probability in fixed rate SW coding is interestingly related to that in fixed rate classic source coding with side information available to both the encoder and the decoder. Specifically, it is eloquently argued in [6] that at rates within a certain range, there exists at least one fixed rate SW code performs (in terms of error exponent) as good as the best classic fixed rate code using the side information at the encoder; and at rates higher than the certain range, the error exponent of fixed rate SW coding is unknown and might be inferior to that of classic fixed rate coding. Similarly, it is not hard to see that when ϵ_n is in a certain range, $R_F^n(\epsilon_n)$ is the same as the redundancy of fixed rate classic source coding with side information available at the encoder.

[2]As $n\to\infty$, ϵ_n will go to 0 fast enough, but not exponentially.

II. Intrinsic Conditional Entropy

Motivated and inspired by [17], in this section, we introduce and analyze the concept of intrinsic entropy, which will play a fundamental role in our performance analysis of SW coding in terms of the tradeoff between the redundancy and decoding error probability.

We first describe the notation to be used throughout this paper. Let $\mathcal{A}=\{a_1,\ldots,a_m\}$ be a finite set. The notation $|\mathcal{A}|$ stands for the cardinality of $\mathcal{A}$, and for any finite sequence x from $\mathcal{A}$, $|x|$ denotes the length of x. For any positive integer n, $\mathcal{A}^n$ denotes the set of all sequences of length n from $\mathcal{A}$. For convenience, we will sometimes write $x_mx_{m+1}\cdots x_n$ as x_m^n, where $m\le n$ are two integers, or simply as x^n when $m=1$. A similar convention will be applied to sequences of random variables as well. We use $\mathcal{P}(\mathcal{A})$ to denote the set of all probability distributions on $\mathcal{A}$, and $\mathcal{P}^+(\mathcal{A})$ to denote the subset of $\mathcal{P}(\mathcal{A})$ where probability distributions with zero entries are excluded. Let π denote a probability distribution in $\mathcal{P}(\mathcal{A}\times\mathcal{B})$. The marginal distributions of π over $\mathcal{A}$ and $\mathcal{B}$ are referred to as $\pi_{\mathcal{A}}$ and $\pi_{\mathcal{B}}$, respectively. The conditional distribution $\pi_{\mathcal{A}|\mathcal{B}}$ is defined by

$$\pi_{\mathcal{A}|\mathcal{B}}(x|y)\triangleq\begin{cases}\frac{\pi(x,y)}{\pi_{\mathcal{B}}(y)}, & \text{for } (x,y)\in\mathcal{A}\times\mathcal{B} \text{ when } \pi_{\mathcal{B}}(y)>0\\ \frac{1}{|\mathcal{A}|}, & \text{for } (x,y)\in\mathcal{A}\times\mathcal{B} \text{ when } \pi_{\mathcal{B}}(y)=0.\end{cases}$$

Occasionally, we shall also write $\pi(x,y)$ as $\pi_{\mathcal{B}}\circ\pi_{\mathcal{A}|\mathcal{B}}$ for convenience. Unless specified otherwise, log denotes the logarithm to base 2, ln denotes the natural logarithm, and e denotes the base of the natural logarithm.

Our analysis in this paper makes heavy use of the method of types [4]. An m-tuple

$$t=(t(a_1),\ldots,t(a_m))\in\mathcal{P}(\mathcal{A})$$

is said to be an n-type if for any $a\in\mathcal{A}$, $t(a)\in\{0,1/n,2/n,\ldots,1\}$. The set of all n-types on $\mathcal{A}$ is denoted by $\mathcal{T}_n(\mathcal{A})$. The type of a sequence $x^n\in\mathcal{A}^n$ is defined as $\tau(x^n)\triangleq(\tau(x^n,a_1),\ldots,\tau(x^n,a_m))$ which is an n-type on $\mathcal{A}$, where $\tau(x^n,a_i)\triangleq\frac{|\{j:\,x_j=a_i\}|}{n}$. For $t\in\mathcal{T}_n(\mathcal{A})$, $T_{\mathcal{A}}^n(t)$ denotes the set of all length-n sequences from $\mathcal{A}$ with type t, i.e., $T_{\mathcal{A}}^n(t)\triangleq\{x^n\in\mathcal{A}^n:\tau(x^n)=t\}$.

We now introduce the notion of *intrinsic entropy*. For $\pi\in\mathcal{P}(\mathcal{A}\times\mathcal{B})$ and $\delta\ge0$, we define the intrinsic δ-entropy of π as

$$H_{\text{in}}(\pi,\delta)\triangleq\sup_{\hat\pi\in\mathcal{P}(\mathcal{A}\times\mathcal{B}):D(\hat\pi\|\pi)\le\delta}H(\hat\pi). \tag{2.1}$$

Throughout this paper, $D(p\|q)$ denotes the relative entropy between two distributions, i.e.

$$D(p\|q)\triangleq\sum_{x\in\mathcal{A}}p(x)\log\frac{p(x)}{q(x)}$$

if both p and q are defined over $\mathcal{A}$. Observe that the set

$$\{\hat\pi\in\mathcal{P}(\mathcal{A}\times\mathcal{B}):\ D(\hat\pi\|\pi)\le\delta\}$$

is convex and compact. Since the entropy function is continuous, the supremum in (2.1) is attainable.

From (2.1), it is easy to see that

$$H(\pi) \le H_{\text{in}}(\pi,\delta) \le \log|\mathcal{A}| + \log|\mathcal{B}|.$$

For any $\pi \in \mathcal{P}^+(\mathcal{A}\times\mathcal{B})$, let $\delta_{\max}(\pi) \triangleq \min\{\delta : H_{\text{in}}(\pi,\delta) = \log|\mathcal{A}| + \log|\mathcal{B}|\}$. We then have the following result.

Lemma 1: The intrinsic entropy $H_{\text{in}}(\pi,\delta)$ has the following properties:

1) $H_{\text{in}}(\pi,\delta)$ is a concave function of (π,δ);
2) $H_{\text{in}}(\pi,\delta)$ is continuous in $\mathcal{P}^+(\mathcal{A}\times\mathcal{B})\times(0,\infty)$; and
3) for any fixed $\pi \in \mathcal{P}^+(\mathcal{A}\times\mathcal{B})$, $H_{\text{in}}(\pi,\delta)$ is a strictly increasing function in $[0,\delta_{\max}(\pi)]$.

The proof of Lemma 1 folllows immediately from the definition of $H_{\text{in}}(\pi,\delta)$ in (2.1).

Extending the notion of intrinsic entropy, we define the intrinsic δ-conditional entropy of π as follows:

$$H_{\text{in}|\mathcal{B}}(\pi,\delta) \triangleq \max_{\hat\pi\in\mathcal{P}(\mathcal{A}\times\mathcal{B}):D(\hat\pi||\pi)\le\delta} [H(\hat\pi) - H(\hat\pi_{\mathcal{B}})] \quad (2.2)$$

where $\hat\pi_{\mathcal{B}}$ denotes the marginal of $\hat\pi$ over $\mathcal{B}$. Note that

$$H(\hat\pi) - H(\hat\pi_{\mathcal{B}}) = \sum_{y\in\mathcal{B}} \hat\pi_{\mathcal{B}}(y) H(\hat\pi_{\mathcal{A}|\mathcal{B}}(\cdot|y)).$$

Similarly, the intrinsic δ-conditional entropy of π with a constrained marginal $\pi_{\mathcal{A}}$ is defined by

$$H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}}) \triangleq \max_{\hat\pi\in\mathcal{P}(\mathcal{A}\times\mathcal{B}):D(\hat\pi||\pi)\le\delta,\pi_{\mathcal{A}}=\hat\pi_{\mathcal{A}}} [H(\hat\pi) - H(\hat\pi_{\mathcal{B}})]. \quad (2.3)$$

The definitions (2.2) and (2.3) will be used later to analyze the redundancy of fixed rate and variable rate SW coding, respectively.

From (2.2) and (2.3), it is easy to see that $H(\pi) - H(\pi_{\mathcal{B}}) \le H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}}) \le H(\pi_{\mathcal{A}})$, and $H(\pi) - H(\pi_{\mathcal{B}}) \le H_{\text{in}|\mathcal{B}}(\pi,\delta) \le \log|\mathcal{A}|$. For any $\pi \in \mathcal{P}^+(\mathcal{A}\times\mathcal{B})$, let $\delta_{\max}(\pi|\pi_{\mathcal{A}}) \triangleq \min\{\delta : H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}}) = H(\pi_{\mathcal{A}})\}$, and $\delta'_{\max}(\pi) \triangleq \min\{\delta : H_{\text{in}|\mathcal{B}}(\pi,\delta) = \log|\mathcal{A}|\}$. We have the following result.

Lemma 2: The intrinsic conditional entropy $H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}})$ ($H_{\text{in}|\mathcal{B}}(\pi,\delta)$, respectively) has the following properties:

1) it is a concave function of (π,δ);
2) it is continuous in $\mathcal{P}^+(\mathcal{A}\times\mathcal{B})\times(0,\infty)$; and
3) for any fixed $\pi \in \mathcal{P}^+(\mathcal{A}\times\mathcal{B})$, it is a strictly increasing function in $[0,\delta_{\max}(\pi|\pi_{\mathcal{A}})]$ ($[0,\delta'_{\max}(\pi)]$, respectively).

Proof of Lemma 2: We briefly show Property 1) for $H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}})$ by proving

$$H_{\text{in}|\mathcal{B}}(\pi^*,\lambda\delta+(1-\lambda)\delta'|\pi^*_{\mathcal{A}}) \ge \lambda H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}}) + (1-\lambda) H_{\text{in}|\mathcal{B}}(\pi',\delta'|\pi'_{\mathcal{A}})$$

where $\pi^* \triangleq \lambda\pi + (1-\lambda)\pi'$, and $0 \le \lambda \le 1$. Let $\hat\pi, \hat\pi' \in \mathcal{P}(\mathcal{A}\times\mathcal{B})$ be two distributions such that

i) $D(\hat\pi||\pi) \le \delta$ and $H(\hat\pi) = H_{\text{in}|\mathcal{B}}(\pi,\delta|\pi_{\mathcal{A}})$; and
ii) $D(\hat\pi'||\pi') \le \delta'$ and $H(\hat\pi') = H_{\text{in}|\mathcal{B}}(\pi',\delta'|\pi'_{\mathcal{A}})$.

For brevity, let $\hat\pi^* \triangleq \lambda\hat\pi + (1-\lambda)\hat\pi'$. From the convexity of relative entropy, we have that

$$\begin{aligned} D(\hat\pi^*||\pi^*) &\le \lambda D(\hat\pi||\pi) + (1-\lambda) D(\hat\pi'||\pi') \\ &\le \lambda\delta + (1-\lambda)\delta'. \end{aligned} \quad (2.4)$$

Note that if $\hat\pi_{\mathcal{A}} = \pi_{\mathcal{A}}$ and $\hat\pi'_{\mathcal{A}} = \pi'_{\mathcal{A}}$, then $\hat\pi^*_{\mathcal{A}} = \pi^*_{\mathcal{A}}$. We get

$$\begin{aligned} &H(\hat\pi^*) - H(\hat\pi^*_{\mathcal{B}}) - \lambda[H(\hat\pi) \\ &\quad - H(\hat\pi_{\mathcal{B}})] - (1-\lambda)[H(\hat\pi') - H(\hat\pi'_{\mathcal{B}})] \\ &\quad = \lambda D(\hat\pi||\hat\pi^*) + (1-\lambda) D(\hat\pi'||\hat\pi^*) \\ &\qquad - \lambda D(\hat\pi_{\mathcal{B}}||\hat\pi^*_{\mathcal{B}}) - (1-\lambda) D(\hat\pi'_{\mathcal{B}}||\hat\pi^*_{\mathcal{B}}) \\ &\quad \ge 0. \end{aligned} \quad (2.5)$$

In the above, the first equality is due to

$$\begin{aligned} H(\hat\pi^*) &= \sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} \left[\lambda\hat\pi(x,y)\log\frac{1}{\hat\pi^*(x,y)} \right. \\ &\qquad \left. +(1-\lambda)\hat\pi'(x,y)\log\frac{1}{\hat\pi^*(x,y)}\right] \\ &= \lambda[H(\hat\pi) + D(\hat\pi||\hat\pi^*)] \\ &\quad + (1-\lambda)[H(\hat\pi') + D(\hat\pi'||\hat\pi^*)] \end{aligned}$$

and a similar expansion of $H(\hat\pi^*_{\mathcal{B}})$; and the last inequality follows from applying the log-sum inequality to $D(\hat\pi||\hat\pi^*)$ and $D(\hat\pi'||\hat\pi^*)$. Inequalities (2.4) and (2.5) together imply Property 1). Similarly, we can prove Property 1) for $H_{\text{in}|\mathcal{B}}(\pi,\delta)$. Properties 2) and 3) follow immediately from (2.2) and (2.3) and Property 1). This completes the proof of Lemma 2. □

In order to gain insights into intrinsic entropy and intrinsic conditional entropy, and more importantly, to see how they can be used to analyze the redundancy in source coding, we relate them to classical conditional entropy. Let (X,Y) be a pair of random variables with joint distribution P_{XY} and alphabet $\mathcal{X}\times\mathcal{Y}$. Let P_X and P_Y denote the marginals of P_{XY} over $\mathcal{X}$ and $\mathcal{Y}$, respectively; and let $P_{Y|X}$ and $P_{X|Y}$ denote the conditional probability distributions of Y given X and X given Y, respectively. The following two lemmas are proved in Appendices A and B, respectively.

Lemma 3 (Intrinsic Versus Classical): Suppose $I(X;Y) > 0$. There exists a constant $d > 0$ such that

$$H_{\text{in}}(P_{XY},\delta) \ge H(P_{XY}) + d\sqrt{\delta}$$

for all δ less than some threshold.

Lemma 4 (Intrinsic Conditional vs Classical Conditional): Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X}\times\mathcal{Y})$. Then the following results hold:

- If $I(X;Y) > 0$ or X is not uniformly distributed over $\mathcal{X}$, then there exists a $\Delta_1 > 0$ such that for any $\delta \le \Delta_1$

$$H_{\text{in}|\mathcal{Y}}(P_{XY},\delta) = H(X|Y) + d_f\sqrt{\delta} + O(\delta)$$

where

$$d_f = \sqrt{2\ln 2}\left[\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y) \log^2\frac{P_Y(y)}{P_{XY}(x,y)} - H^2(X|Y)\right]^{\frac{1}{2}}. \quad (2.6)$$

- If $I(X;Y)>0$, then for any $0<\beta<1/|\mathcal{X}|$, there exists a $\Delta_2>0$ depending only on β and $P_{Y|X}$ such that for any $\delta\le\Delta_2$ and any distribution $t\in\mathcal{P}(\mathcal{X})$ satisfying $t(x)\ge\beta$ for all $x\in\mathcal{X}$

$$H_{\text{in}|\mathcal{Y}}(t\circ P_{Y|X},\delta|t) = H(t\circ P_{Y|X}) - H(r) + d_v(t)\sqrt{\delta} + O(\delta)$$

where $r=(t\circ P_{Y|X})_{\mathcal{Y}}$ and

$$d_v(t) = \sqrt{(2\ln 2)} \times \left[\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} t(x)P_{Y|X}(y|x)\log^2\frac{r(y)}{P_{Y|X}(y|x)} - \sum_{x\in\mathcal{X}} t(x)D^2(P_{Y|X}(\cdot|x)||r)\right]^{\frac{1}{2}}. \quad (2.7)$$

Remark 1:
Lemma 3 can be strengthened to have an expression similar to that of Lemma 4.

Remark 2: The quantity inside the bracket of the equation defining d_f can be interpreted as the variance of $\log P_{X|Y}(X|Y)$ (taken with respect to P_{XY}). Similarly, the quantity inside the bracket of the equation defining $d_v(t)$ can be interpreted as the average (over $t(x)$) of the variance of $\log\frac{r(Y)}{P_{Y|X}(Y|x)}$ (taken with respect to $P_{Y|X}$). The conditions on P_{XY} assumed in Lemma 4 above are made so that these two variances are bounded away from 0, respectively.

In the following, we will make use of intrinsic entropy, intrinsic conditional entropy, and the above lemmas to investigate the performance of SW coding of X with decoder side information Y in terms of the tradeoff between the redundancy and decoding error probability.

III. Lower Bounds

Let (X,Y) be a memoryless source-side information pair with finite alphabet $\mathcal{X}\times\mathcal{Y}$. In this section, we establish a lower bound on the compression rate of SW coding of X with decoder side information Y and a given decoding error probability.

We begin with the formal definition of variable rate SW coding. Let $\mathcal{I}$ denote a set of finite binary codewords satisfying the prefix condition. An order n SW code C_n is described by a pair $C_n=(f_n,g_n)$, where $f_n(\cdot):\mathcal{X}^n\to\mathcal{I}$, acting as an encoder, maps source sequences of block length n from $\mathcal{X}$ to binary codewords in $\mathcal{I}$, and $g_n(\cdot,\cdot):\mathcal{I}\times\mathcal{Y}^n\to\mathcal{X}^n$, acting as a decoder, reconstructs the encoded source sequences upon receiving codewords and with the help of the side information sequences. Since the mapping f_n is often many-to-one, we sometimes refer to an entry $b\in\mathcal{I}$ as a bin index as in the literature of SW coding. The order n SW code C_n is called a fixed rate code if $\mathcal{I}$ consists of binary codewords of the same length, and a variable rate code if codewords may have different lengths. Clearly, the class of all (order n) variable rate SW codes includes that of all (order n) fixed rate SW codes as a strict subclass.

When $C_n=(f_n,g_n)$ is applied to encode X^n, the resulting average compression rate $r(C_n)$ is given by

$$r(C_n)\triangleq\frac{1}{n}\mathrm{E}|f_n(X^n)|$$

where $|z|$ denotes the length in bits of the binary sequence z. On the decoder side, let $\hat{X}^n=g_n(f_n(X^n),Y^n)$ denote the decoder output. The decoding error probability of C_n is given by

$$P_e(C_n)\triangleq\Pr\{X^n\ne\hat{X}^n\}.$$

In [13], it was shown that $r(C_n)$ can be made arbitrarily close to $H(X|Y)$ while maintaining a small decoding error probability $P_e(C_n)$ when n is large enough. On the other hand, it is clear that the smaller $P_e(C_n)$, the larger $r(C_n)$. Therefore, to fully understand the performance of SW coding, it is desirable to investigate the best tradeoff between $r(C_n)$ and $P_e(C_n)$ among all order n SW codes C_n. Under the condition that $P_e(C_n)\le\epsilon_n$, a prescribed threshold, we first derive lower bounds on $r(C_n)$ in Theorems 1 and 2 below for variable rate and fixed rate SW coding, respectively. The proofs of these two theorems are deferred to Section VI.

Theorem 1: Assume $P_{XY}\in\mathcal{P}^+(\mathcal{X}\times\mathcal{Y})$ and $I(X;Y)>0$. Let $\{\epsilon_n\}$ be a sequence of positive real numbers satisfying $\epsilon_n=o(\sqrt{\frac{\log n}{n}})$ and $-\log\epsilon_n=o(n)$. Then for sufficiently large n and any order n variable rate code $C_n=(f_n,g_n)$ with

$$P_e(C_n)=\Pr\{X^n\ne\hat{X}^n\}\le\epsilon_n \quad (3.1)$$

one has

$$r(C_n)\ge H(X|Y)+d_v\sqrt{\frac{-\log\epsilon_n}{n}} + O\left(\frac{-\log\epsilon_n}{n}\right)-o\left(\sqrt{\frac{\log n}{n}}\right) \quad (3.2)$$

where

$$d_v=\sqrt{(2\ln 2)}\left[\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}P_{XY}(x,y)\log^2\frac{P_Y(y)}{P_{Y|X}(y|x)} - \sum_{x\in\mathcal{X}}P_X(x)D^2(P_{Y|X}(\cdot|x)||P_Y)\right]^{1/2}. \quad (3.3)$$

Theorem 2: Assume $P_{XY}\in\mathcal{P}^+(\mathcal{X}\times\mathcal{Y})$. Let $\Delta\triangleq\delta'_{\max}(P_{XY})$ be defined as in Section II. Then for sufficiently large n, the following hold.

a) For any order n fixed rate code $C_n = (f_n, g_n)$ with decoding error probability $P_e(C_n) = \epsilon_n$,

$$r(C_n) \geq H_{\text{in}|\mathcal{Y}}\left(P_{XY}, \frac{-\log \epsilon_n}{n}\right) - O\left(\frac{1}{\sqrt{n}}\right)$$

whenever $d_1 2^{-\Delta n} \leq \epsilon_n \leq d_2$, where $d_1 > 1$ and $d_2 < 1$ are constants depending only on P_{XY}.

b) If $I(X;Y) > 0$ or X is not uniformly distributed over $\mathcal{X}$, then for any order n fixed rate code C_n with decoding error probability $P_e(C_n) = \epsilon_n$,

$$r(C_n) \geq H(X|Y) + d_f\sqrt{\frac{-\log \epsilon_n}{n}} + O\left(\frac{-\log \epsilon_n}{n}\right) - O\left(\frac{1}{\sqrt{n}}\right)$$

whenever $d_1 2^{-\Delta_1 n} \leq \epsilon_n \leq d_2$, where Δ_1 is specified in Part 1 of Lemma 4, and d_f is given by (2.6).

Remark 3: We say that the above coding theorem is proved from the encoder's perspective because the center of argument is placed on the encoder, where we are mainly concerned that under a probability constraint, the encoder can pack how many distinct source sequences into one bin. The careful reader might have found out that this type of packing argument is similar to the sphere packing method in channel coding. Indeed, our proof above clearly demonstrates that the underlying connection between SW coding and channel coding: each bin performs like a channel code where every sequence in the same bin is mapped to a set of side information sequences equivalent to channel realizations.

The coding theorem above can also be proven from the decoder's perspective [9]. From this perspective, the center of argument is placed on the decoder, where we are mainly concerned that under a probability constraint, the decoder has to receive enough bits to decode a sufficient number of distinct source sequences. As a result, the argument relies more on the classical Kraft's inequality as in the analysis of classical lossless source codes. From this perspective, the connection between SW coding and classical source coding becomes apparent: for each typical side information sequence, one can extract an good embedded source code from a good SW code.

Remark 4: The assumption that $\epsilon_n = o(\sqrt{\frac{\log n}{n}})$ has intuitive explanations in addition to being viewed as the logical consequence of our technical argument above. When $\epsilon_n = \omega(\sqrt{\frac{\log n}{n}})$, one can construct a code (see Remark 7 in Section IV) that achieves the conditional entropy $H(X|Y)$ from below by discarding some atypical source types whose probability is in the order of $\omega(\sqrt{\frac{\log n}{n}})$, and the reduction in rates cannot be offset by the redundancy needed for encoding the typical types. Thus, in contrast to fixed rate coding, there is no $O(1/\sqrt{n})$ redundancy in variable rate SW coding. This clearly demonstrates the fundamental difference between fixed rate and variable rate SW coding in the regime of slow decaying error probabilities.

Remark 5: Theorems 1 and 2 can be extended to the case where $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ satisfying the following condition.

C1: $H(X|Y)$ is strictly less than the minimum zero-error coding rate of X achievable asymptotically with decoder only side information Y [18], [1], [10].

In [1], [10], the minimum zero-error coding rate of X achievable asymptotically with decoder only side information Y is shown to be given by the *complementary graph entropy* $\bar{H}(\mathcal{G}, X)$, where $\mathcal{G}$ denotes the *characteristic graph* associated with (X,Y) [18]. Together with the fact that $H(X) \geq \bar{H}(\mathcal{G}, X)$, Condition C1 implies that

$$I(X;Y) \geq \bar{H}(\mathcal{G}, X) - H(X|Y) > 0.$$

To extend the two theorems to $P_{XY} \notin \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$ satisfying Condition C1, one can follow an argument similar to that used to prove Theorem 1 in Section VI. Some necessary modifications are in order, which begin with the definitions of the intrinsic δ-conditional entropy, $H_{\text{in}|\mathcal{B}}(\pi, \delta)$, and the intrinsic δ-conditional entropy with a constrained marginal $\pi_{\mathcal{A}}$, $H_{\text{in}|\mathcal{B}}(\pi, \delta|\pi_{\mathcal{A}})$, in (2.2) and (2.3), respectively. Specifically, the maximum in (2.2) and (2.3) is now taken over all $\hat{\pi}$ whose support set [1] is a subset of the support set of π. The rest of the changes follow accordingly, and the details are omitted as they do not provide much new insight.

IV. Upper Bounds

In this section, the lower bounds established in Section III are shown to be tight by establishing the respective matching upper bounds. Specifically, we have the following two theorems.

Theorem 3: Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$ and $I(X;Y) > 0$. Let $\{\epsilon_n\}$ be a sequence of positive real numbers such that $-\log \epsilon_n = o(n)$ and $\lim_{n\to\infty} \epsilon_n = 0$. Then there exists a sequence of variable rate SW codes $\{C_n\}_{n=1}^{\infty}$ with $P_e(C_n) \leq \epsilon_n$ such that for sufficiently large n

$$r(C_n) \leq H(X|Y) + d_v\sqrt{\frac{-\log \epsilon_n}{n}} + o\left(\sqrt{\frac{-\log \epsilon_n}{n}}\right) \quad (4.1)$$

where d_v is given by (3.3).

Theorem 4: Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$. Further assume either $I(X;Y) > 0$ or X is not uniformly distributed. Let $\{\epsilon_n\}$ be a sequence of positive real numbers such that $-\log \epsilon_n = o(n)$ and $\lim_{n\to\infty} \epsilon_n = 0$. Then there exists a sequence of fixed rate codes $\{C_n\}_{n=1}^{\infty}$ with $P_e(C_n) \leq \epsilon_n$ such that for sufficiently large n,

$$r(C_n) = H(X|Y) + d_f\sqrt{\frac{-\log \epsilon_n}{n}} + o\left(\sqrt{\frac{-\log \epsilon_n}{n}}\right) \quad (4.2)$$

where d_f is given by (2.6).

The proofs of Theorem 3 and 4 are provided in Section VII.

Remark 6: Combining the proof of Theorem 1 with that of Theorem 3, we see that an efficient variable rate SW code should assign the same average decoding error probability to each bin,

and the same decoding error probability to each sequence with the same type.

Remark 7: In view of Theorem 3 and its proof in Section VII, we see that for any constant ϵ (not depending upon n), there exists a sequence of variable rate SW codes $\{C_n\}_{n=1}^{\infty}$ such that for sufficiently large n,

$$P_e(C_n) \le \epsilon$$
$$r(C_n) \le (1-\delta)H(X|Y) + O\left(\sqrt{\frac{\log n}{n}}\right) \qquad (4.3)$$

where $0 < \delta < \epsilon$ is a constant. Briefly, the construction of C_n is similar to that in the proof of Theorem 3 with the following modifications. Let x^n be the sequence to be encoded with type t. In Step 2 on the encoder side, if $||t - P_X||_1 \ge c_0$ where c_0 is a constant selected so that

$$\delta \le \Pr\{||\tau(X^n) - P_X||_1 \ge c_0\} < \epsilon$$

the encoder sends nothing to the decoder in this step; otherwise, select $R(t)$ so that

$$\Pr\{X^n \ne \hat{X}^n | \tau(X^n) = t\} \le \frac{1}{n^2}$$

and perform random binning of x^n with $2^{nR(t)}$ bins. Step 2 on the decoder side is modified accordingly to reconstruct an arbitrary sequence whenever $||t - P_X||_1 \ge c_0$. An argument similar to that used in the proof of Theorem 3 can then be used to prove (4.3). Further generalizing the construction of C_n, one sees there exist SW codes $\{C_n\}$ with $P_e(C_n) = \omega(\sqrt{\frac{\log n}{n}})$ that achieve $H(X|Y)$ from below, echoing Remark 4.

Remark 8: It is well known that in classical lossless coding, the redundancy of zero-error variable rate coding, which is in the order $O(1/n)$, is much better than that of fixed rate coding, which is in the order $\Theta(\sqrt{\frac{-\log \epsilon_n}{n}})$, where ϵ_n denotes the decoding error probability. Though one might expect a similar result in SW coding, Theorem 1 in Section III and Theorem 4 paint a different picture: for a wide range of error probabilities, the redundancy of variable rate SW coding and that of fixed rate SW coding are in fact in the same order $\Theta(\sqrt{\frac{-\log \epsilon_n}{n}})$. Since the redundancy of variable rate SW coding was never fully characterized before, this result, to the best knowledge of the authors, is the first of its kind in the literature showing that one cannot distinguish variable rate SW coding from fixed rate SW coding simply by their respective redundancy order. In order to demonstrate that variable rate SW coding is indeed more efficient than fixed rate SW coding in general, in the next section we shall take a closer look at the constant terms d_f and d_v.

V. Variable Rate vs Fixed Rate

Having established tight performance upper and lower bounds for both variable rate and fixed rate SW coding, we are now in a position to compare the performance of variable rate SW coding with that of fixed rate SW coding for large finite block lengths.

Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$ and $I(X;Y) > 0$. Let $\{\epsilon_n\}$ be a sequence of positive real numbers satisfying

$$\epsilon_n = o\left(\sqrt{\frac{\log n}{n}}\right)$$

and $-\log \epsilon_n = o(n)$. From Sections III and IV, it follows that for large n, the best compression performance in bits per symbol of order n variable rate SW coding with the decoding error probability less than or equal to ϵ_n and that of order n fixed rate SW coding with the decoding error probability $\le \epsilon_n$ are equal to

$$H(X|Y) + d_v\sqrt{\frac{-\log \epsilon_n}{n}} + o\left(\sqrt{\frac{-\log \epsilon_n}{n}}\right) \text{ and}$$
$$H(X|Y) + d_f\sqrt{\frac{-\log \epsilon_n}{n}} + o\left(\sqrt{\frac{-\log \epsilon_n}{n}}\right)$$

respectively. Thus the comparison between them for large finite block lengths n boils down to comparing d_v with d_f. To this end, we have the following result.

Theorem 5: Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$. Then we have

$$d_v \le d_f$$

with equality if and only if

$$-\log P_X(x) - D(P_{Y|X}(\cdot|x)||P_Y)$$

is a constant, i.e., does not depend on the actual symbol x.

Proof of Theorem 5: In view of the definitions of d_v and d_f in (3.3) and (2.6), respectively, we see that in order to prove Theorem 5, it suffices to compare $d_v^2/(2\ln 2)$ against $d_f^2/(2\ln 2)$. It is not hard to verify that

$$\begin{aligned}
&\left(d_f^2 - d_v^2\right)/(2\ln 2) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\left[\log^2\frac{P_Y(y)}{P_{XY}(x,y)} - \log^2\frac{P_Y(y)}{P_{Y|X}(y|x)}\right] - H^2(X|Y) \\
&\quad + \sum_{x\in\mathcal{X}} P_X(x) D^2(P_{Y|X}(\cdot|x)||P_Y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\left[\log\left(\frac{P_Y(y)}{P_{XY}(x,y)}\frac{P_Y(y)}{P_{Y|X}(y|x)}\right) \times \log\left(\frac{1}{P_X(x)}\right)\right] - H^2(X|Y) \\
&\quad + \sum_{x\in\mathcal{X}} P_X(x) D^2(P_{Y|X}(\cdot|x)||P_Y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\left[\left\{\log\left(\frac{P_Y(y)}{P_{Y|X}(y|x)}\right)^2 + \log\left(\frac{1}{P_X(x)}\right)\right\}\log\left(\frac{1}{P_X(x)}\right)\right] - H^2(X|Y) \\
&\quad + \sum_{x\in\mathcal{X}} P_X(x) D^2(P_{Y|X}(\cdot|x)||P_Y) \\
&= \sum_{x\in\mathcal{X}} P_X(x)\left[\log^2 P_X(x)\right.
\end{aligned}$$

$$
\begin{aligned}
&-2D(P_{Y|X}(\cdot|x)||P_Y)\log\left(\frac{1}{P_X(x)}\right)\Bigg] - H^2(X|Y)\\
&+\sum_{x\in\mathcal{X}} P_X(x)D^2(P_{Y|X}(\cdot|x)||P_Y)\\
&=\sum_{x\in\mathcal{X}} P_X(x)\left[\log\left(\frac{1}{P_X(x)}\right)\right.\\
&\left.-D(P_{Y|X}(\cdot|x)||P_Y)\right]^2 - H^2(X|Y)\\
&\overset{1)}{\geq}\left[\sum_{x\in\mathcal{X}} P_X(x)\left\{\log\left(\frac{1}{P_X(x)}\right)\right.\right.\\
&\left.\left.-D(P_{Y|X}(\cdot|x)||P_Y)\right\}\right]^2 - H^2(X|Y)\\
&=0.
\end{aligned}
$$

In the above, the inequality 1) is due to the nonnegativity of variance with equality if and only if

$$\log\left(\frac{1}{P_X(x)}\right) - D(P_{Y|X}(\cdot|x)||P_Y)$$

is a constant. This completes the proof of Theorem 5. □

Remark 9: Theorem 5 implies that even though variable rate SW coding and fixed rate SW coding approach asymptotically the same compression limit $H(X|Y)$ at a speed of the same order $O(\sqrt{\frac{-\log\epsilon_n}{n}})$, for finite block lengths n, variable rate SW coding is indeed more efficient than fixed rate SW coding in general.

Remark 10: It is also interesting and surprising to see that for sources for which $-\log P_X(x) - D(P_{Y|X}(\cdot|x)||P_Y)$ is a constant, variable rate SW coding and fixed rate SW coding have the same compression performance up to the second order inclusive. This implies that for these types of sources, there is no need to use variable rate SW coding in practice.

We conclude this section with an example to demonstrate the difference between d_v and d_f.

Example: Consider the case where X takes values in the binary alphabet $\{0,1\}$ with $\Pr\{X=0\}=p$, and the channel from X to Y is a binary symmetric channel (BSC) with crossover probability q i.e.

$$\Pr\{Y=y|X=x\} = \begin{cases} q, & \text{if } y\neq x\\ 1-q, & \text{if } y=x\end{cases} \tag{5.1}$$

We will assume that $0<p,q\leq\frac{1}{2}$. When $p=0.5$,[3]

$$d_v = d_f = \sqrt{(2\ln 2)}\left(\log\frac{1-q}{q}\right)\sqrt{q(1-q)}.$$

When $p\neq 0.5$, $d_v<d_f$; Fig. 1 shows the curves of d_v and d_f as a function of $q\in(0,\frac{1}{2})$ for $p=0.05, 0.25, 0.45$.

[3]The redundancy of fixed rate SW coding for the case where $p=0.5$ was considered in [2, Theorem 2], where the term $(\log\frac{1-q}{q})\sqrt{q(1-q)}$ is obtained by using an argument based on the central limit theorem.

VI. Proof of Theorems 1 and 2

In this section, we prove Theorems 1 and 2.

Proof of Theorem 1: Let $C_n=(f_n,g_n)$ be an order n variable rate code with

$$
\begin{aligned}
P_e(C_n) &= \Pr\{X^n\neq\hat{X}^n\}\\
&= \sum_{x^n\in\mathcal{X}^n}\Pr\{X^n=x^n\}\epsilon_{x^n}\leq\epsilon_n
\end{aligned}\tag{6.2}
$$

where $\epsilon_{x^n}\triangleq\Pr\{\hat{X}^n\neq X^n|X^n=x^n\}$. Let α be a positive number to be specified later. Define

$$\Phi(\alpha)\triangleq\left\{x^n\in\mathcal{X}^n:\epsilon_{x^n}\leq\alpha, ||\tau(x^n)-P_X||_1\leq c_0\sqrt{\frac{\log n}{n}}\right\}$$

where the constant c_0 is chosen so that

$$\Pr\left\{||\tau(X^n)-P_X||_1>c_0\sqrt{\frac{\log n}{n}}\right\}\leq n^{-2}. \tag{6.3}$$

Throughout the paper, $||\cdot||_1$ denotes the L1 distance. For brevity, let $\Gamma_{\mathcal{X}}$ denote the subset of $\mathcal{T}_n(\mathcal{X})$ such that

$$\Gamma_{\mathcal{X}}\triangleq\left\{t\in\mathcal{T}_n(\mathcal{X}):||t-P_X||_1\leq c_0\sqrt{\frac{\log n}{n}}\right\}.$$

From Markov's inequality, it follows that $\Pr\{\epsilon_{X^n}>\alpha\}\leq\epsilon_n/\alpha$, which, together with (6.3), implies

$$\Pr\{X^n\notin\Phi(\alpha)\}\leq\frac{\epsilon_n}{\alpha}+\frac{1}{n^2}. \tag{6.4}$$

Define a binary random variable A such that A is equal to 1 if $X^n\in\Phi(\alpha)$ and 0 otherwise. Furthermore, for any $b\in\mathcal{I}$ and $t\in\mathcal{T}_n(\mathcal{X})$, define

$$
\begin{aligned}
p(t)&\triangleq\Pr\{\tau(X^n)=t\}\\
p(1,t)&\triangleq\Pr\{A=1,\tau(X^n)=t\}\\
p(b,1,t)&\triangleq\Pr\{f_n(X^n)=b,A=1,\tau(X^n)=t\}\\
p(1|t)&\triangleq\Pr\{A=1|\tau(X^n)=t\}\ \text{and}\\
p(b|1,t)&\triangleq\Pr\{f_n(X^n)=b|A=1,\tau(X^n)=t\}.
\end{aligned}
$$

Since $\mathcal{I}$ is a prefix set, it is easy to see that

$$
\begin{aligned}
nr(C_n)&\geq H(f_n(X^n))\geq H(f_n(X^n)|A,\tau(X^n))\\
&\geq\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(1,t)H(f_n(X^n)|A=1,\tau(X^n)=t).
\end{aligned}\tag{6.5}
$$

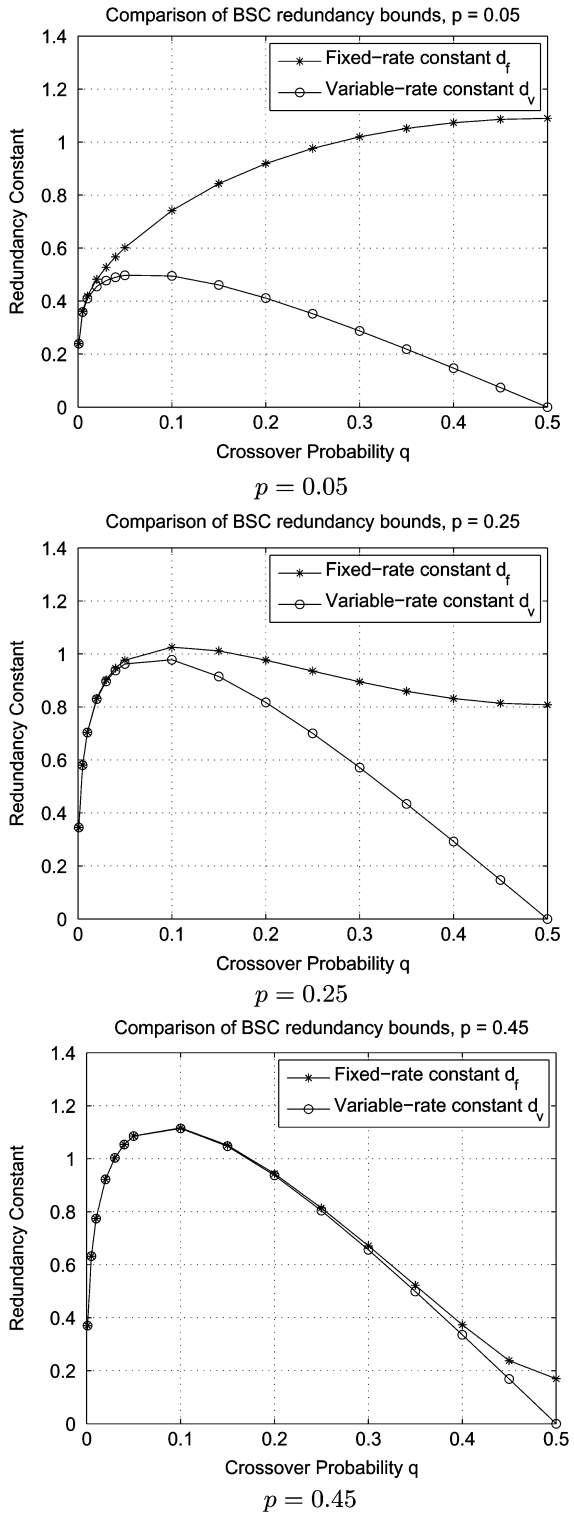

Fig. 1. Comparison of redundancy constants d_f and d_v.

Note that $p(1,t) = 0$ when $t \notin \Gamma_{\mathcal{X}}$. Consequently, we shall consider only the case where $t \in \Gamma_{\mathcal{X}}$. In view of the definition of entropy, we have for any $t \in \Gamma_{\mathcal{X}}$

$$H(f_n(X^n)|A=1,\tau(X^n)=t) = \log p(1|t) + \sum_{b\in\mathcal{I}} p(b|1,t)\log\frac{p(t)}{p(b,1,t)}$$

which, together with (6.5), implies

$$\begin{aligned} nr(C_n) &\geq \sum_{t\in\mathcal{T}_n(\mathcal{X})} p(1,t) \\ &\quad\times\left[\log p(1|t) + \sum_{b\in\mathcal{I}} p(b|1,t)\log\frac{p(t)}{p(b,1,t)}\right] \\ &\geq -1 + \sum_{t\in\mathcal{T}_n(\mathcal{X})} p(1,t)\left[\sum_{b\in\mathcal{I}} p(b|1,t)\log\frac{p(t)}{p(b,1,t)}\right] \end{aligned} \tag{6.6}$$

where the last inequality is due to the fact that

$$\begin{aligned} &\sum_{t\in\mathcal{T}_n(\mathcal{X})} p(1,t)\log p(1|t) \\ &\quad\geq \sum_{t\in\mathcal{T}_n(\mathcal{X})} p(t)[-H(A|\tau(X^n)=t)] \\ &\quad\geq -1. \end{aligned}$$

Since each sequence x^n in $T^n_{\mathcal{X}}(t)$ is equally probable, it follows that in (6.6)

$$\frac{p(t)}{p(b,1,t)} = \frac{|\{x^n\in\mathcal{X}^n : \tau(x^n)=t\}|}{|\{x^n : f_n(x^n)=b, \tau(x^n)=t, x^n\in\Phi(\alpha)\}|}.$$

Define

$$\begin{aligned} \mathcal{N}_{b,t,\alpha} &\triangleq \{x^n : f_n(x^n)=b \\ &\qquad \tau(x^n)=t, x^n\in\Phi(\alpha)\} \\ \epsilon_\alpha &\triangleq \Pr\{\hat{X}^n\neq X^n | X^n\in\Phi(\alpha)\}, \\ \epsilon_{t,\alpha} &\triangleq \Pr\{\hat{X}^n\neq X^n | \tau(X^n)=t, X^n\in\Phi(\alpha)\}, \text{ and} \\ \epsilon_{b,t,\alpha} &\triangleq \Pr\{\hat{X}^n\neq X^n | \tau(X^n)=t, f_n(X^n)=b \\ &\qquad X^n\in\Phi(\alpha)\}. \end{aligned}$$

We are then led to upper bound $|\mathcal{N}_{b,t,\alpha}|$.

Fix $b\in\mathcal{I}$ and $t\in\mathcal{T}_n(\mathcal{X})$. Since

$$\epsilon_{b,t,\alpha} = \sum_{x^n\in\mathcal{N}_{b,t,\alpha}} \frac{\epsilon_{x^n}}{|\mathcal{N}_{b,t,\alpha}|} \leq \alpha \tag{6.7}$$

we can think of $\mathcal{N}_{b,t,\alpha}$ as a channel code for the memoryless channel given by $P_{Y|X}$ with the average decoding error probability given by $\epsilon_{b,t,\alpha}$. Specifically, for each $x^n\in\mathcal{N}_{b,t,\alpha}$, define

$$G(x^n) \triangleq \{y^n\in\mathcal{Y}^n : g_n(b,y^n)=x^n\}.$$

Then the set mapping from $x^n\in\mathcal{N}_{b,t,\alpha}$ to $G(x^n)$ specifies a channel code for $P_{Y|X}$ with $G(x^n)$ acting as a decoding set in $\mathcal{Y}^n$ for x^n and the average decoding error probability $\epsilon_{b,t,\alpha}$. Since the sets $G(x^n)$, $x^n\in\mathcal{N}_{b,t,\alpha}$, are disjoint, we can use a sphere packing argument to upper bound $|\mathcal{N}_{b,t,\alpha}|$. This is essentially the line we shall follow below.

Applying the sphere packing argument directly to $G(x^n)$, however, encounters two problems: (1) the cardinalities of the sets $G(x^n)$ may not have a tight uniform lower bound, and (2) the cardinality of the union of the sets $G(x^n)$ over all $x^n\in\mathcal{N}_{b,t,\alpha}$ may not have a tight upper bound. To overcome these two

problems, we shall limit our attention to a subset of good "codewords" $x^n \in \mathcal{N}_{b,t,\alpha}$ and the intersection of $G(x^n)$ with some type classes. To this end, define $\mathcal{N}^+_{b,t,\alpha} \triangleq \{x^n \in \mathcal{N}_{b,t,\alpha} : \epsilon_{x^n} \leq 2\epsilon_{b,t,\alpha}\}$. From Markov's inequality, it follows that $|\mathcal{N}_{b,t,\alpha}| \leq 2|\mathcal{N}^+_{b,t,\alpha}|$.[4] For any n-type $s^* \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y})$ with its marginal over $\mathcal{X}$ given by t and any $x^n \in T^n_{\mathcal{X}}(t)$, let

$$B(x^n, s^*) \triangleq \{y^n \in \mathcal{Y}^n : \|\tau(x^n, y^n) - s^*\|_1 \leq \frac{\kappa}{\sqrt{n}}\}$$

where κ is a positive constant. Instead of $G(x^n)$, we shall consider the intersection of $G(x^n)$ with $B(x^n, s^*)$ below for $x^n \in \mathcal{N}^+_{b,t,\alpha}$. In other words, we shall apply the sphere packing argument to the set mapping from $x^n \in \mathcal{N}^+_{b,t,\alpha}$ to $G(x^n) \cap B(x^n, s^*)$.

At this point, we invoke the following lemma, which will be proved in Appendix C, and enables us to tightly lower bound $|G(x^n) \cap B(x^n, s^*)|$.

Lemma 5 ($\frac{1}{\sqrt{n}}$-Neighborhood of a Type): Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X}\times\mathcal{Y})$. Then for any $0 < \beta < 1/(|\mathcal{X}||\mathcal{Y}|)$, there exists a constant $\kappa_0 > 0$ such that for any sufficiently large n, any n-type $t \in \mathcal{T}_n(\mathcal{X})$, any n-type $s^* \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y})$ with $s^*_{\mathcal{X}} = t$ and $s^*(x,y) \geq \beta$, $(x,y) \in \mathcal{X}\times\mathcal{Y}$, and any sequence $x^n \in T^n_{\mathcal{X}}(t)$, the following results hold:

1) $\Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\} \geq 2^{-nD(s^*\|t\circ P_{Y|X}) - \kappa_0}$.
2) $\frac{1}{n}\log|B(x^n, s^*)| \geq H(s^*) - H(t) - \frac{\kappa_0}{n}$.
3) For any $\xi \leq \Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\}$ and any subset $B^+(x^n, s^*)$ of $B(x^n, s^*)$,

$$\begin{aligned}&|B^+(x^n, s^*)|\\ &\quad\geq (\Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\} - \xi)\\ &\qquad\times 2^{n[H(s^*) - H(t) + D(s^*\|t\circ P_{Y|X})] - \kappa_0\sqrt{n}}\end{aligned}$$

whenever

$$\begin{aligned}&\Pr\{Y^n \in B^+(x^n, s^*)|X^n = x^n)\\ &\qquad\geq \Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\} - \xi.\end{aligned}$$

Fix $b \in \mathcal{I}$ and $t \in \Gamma_{\mathcal{X}}$ as we did before Lemma 5. Let κ_1 be a constant to be specified later. In view of Lemma 2, let π^* be a distribution in $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that

$$\begin{aligned}&\pi^*_{\mathcal{X}} = t,\\ &D(\pi^*\|t\circ P_{Y|X}) = \frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n)}{n} - \frac{\kappa_1}{n}, \text{ and}\\ &H(\pi^*) - H(\pi^*_{\mathcal{Y}}) = H_{\text{in}|\mathcal{Y}}\\ &\quad\times\left(t\circ P_{Y|X}, \frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1}{n}|t\right).\end{aligned} \tag{6.8}$$

Note that since $-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1$ increases as n increases, such π^* exists for sufficiently large n. Regarding π^* as a matrix whose row sums equal t, we can construct a type $s^* \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y})$ round π^* as follows.

S1: For every row in π^*, move the first $|\mathcal{Y}| - 1$ elements to the nearest integer multiples of $\frac{1}{n}$.

S2: Set the last column of the resulting matrix so that its row sums equal t.

Verify that $\|s^* - \pi^*\|_1 \leq 2|\mathcal{X}|(|\mathcal{Y}|-1)/n$. It then follows from the standard Taylor's series expansion that for all $t \in \Gamma_{\mathcal{X}}$, there exist constants κ'_1 and κ''_1 such that

$$|D(s^*\|t\circ P_{Y|X}) - D(\pi^*\|t\circ P_{Y|X})| \leq \frac{\kappa'_1}{n} \text{ and} \tag{6.9}$$

$$|H(s^*) - H(s^*_{\mathcal{Y}}) - (H(\pi^*) - H(\pi^*_{\mathcal{Y}}))| \leq \frac{\kappa''_1}{n}. \tag{6.10}$$

Putting (6.8), (6.9), and (6.10) together, we have

$$\begin{aligned}&s^*_{\mathcal{X}} = t,\\ &\left|D(s^*\|t\circ P_{Y|X}) - \frac{-\log(2\epsilon_{b,t,\alpha}+\epsilon_n)-\kappa_1}{n}\right| \leq \frac{\kappa'_1}{n}, \text{ and}\\ &\left|H(s^*) - H(s^*_{\mathcal{Y}}) - H_{\text{in}|\mathcal{Y}}\left(t\circ P_{Y|X}, \frac{-\log(2\epsilon_{b,t,\alpha}\leq+\epsilon_n)-\kappa_1}{n}\middle| t\right)\right| \leq \frac{\kappa''_1}{n}.\end{aligned} \tag{6.11}$$

Select now $\kappa_1 = \kappa'_1 + \kappa_0$. In view of Lemma 5 and its proof, we see that for $x^n \in \mathcal{N}_{b,t,\alpha}$

$$\Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\} \geq 2\epsilon_{b,t,\alpha} + \epsilon_n.$$

For brevity, let us use $B^+(x^n, s^*)$ to denote $G(x^n)\bigcap B(x^n, s^*)$ and $B^-(x^n, s^*)$ to denote the complement of $B^+(x^n, s^*)$ in $B(x^n, s^*)$. Then for $x^n \in \mathcal{N}^+_{b,t,\alpha}$

$$\begin{aligned}&\Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\}\\ &\quad= \Pr\{Y^n \in B^-(x^n, s^*)|X^n = x^n\}\\ &\qquad+ \Pr\{Y^n \in B^+(x^n, s^*)|X^n = x^n\}\\ &\quad\leq \Pr\{Y^n \in B^+(x^n, s^*)|X^n = x^n\} + \epsilon_{x^n}\\ &\quad\leq \Pr\{Y^n \in B^+(x^n, s^*)|X^n = x^n\} + 2\epsilon_{b,t,\alpha}\end{aligned} \tag{6.12}$$

where the last inequality is due to $x^n \in \mathcal{N}^+_{b,t,\alpha}$. Equation (6.12), together with Lemma 5, implies that

$$\begin{aligned}|B^+(x^n, s^*)| &\geq (\Pr\{Y^n \in B(x^n, s^*)|X^n = x^n\} - 2\epsilon_{b,t,\alpha})\\ &\qquad\times 2^{n[H(s^*) - H(t) + D(s^*\|t\circ P_{Y|X})] - \kappa_0\sqrt{n}}\\ &\geq 2^{n[H(s^*) - H(t) + D(s^*\|t\circ P_{Y|X}) + \frac{\log\epsilon_n}{n}] - \kappa_0\sqrt{n}}.\end{aligned} \tag{6.13}$$

Regard $B^+(x^n, s^*)$ as a sphere mapped to $x^n \in \mathcal{N}^+_{b,t,\alpha}$. In order to apply a sphere packing argument as mentioned above to upper bound $\mathcal{N}^+_{b,t,\alpha}$ (and in turn $\mathcal{N}_{b,t,\alpha}$), we need to upper bound

$$\left|\bigcup_{x^n\in\mathcal{N}^+_{b,t,\alpha}} B^+(x^n, s^*)\right| = \sum_{x^n\in\mathcal{N}^+_{b,t,\alpha}} |B^+(x^n, s^*)|$$

[4] Due to Markov's inequality,

$$\Pr\{\epsilon_{X^n} > 2\epsilon_{b,t,\alpha}|X^n \in \mathcal{N}_{b,t,\alpha}\} \leq \frac{\mathbf{E}[\epsilon_{X^n}|X^n \in \mathcal{N}_{b,t,\alpha}]}{2\epsilon_{b,t,\alpha}} = \frac{1}{2}$$

which, together with the fact that all sequences in $\mathcal{N}_{b,t,\alpha}$ are equally probable, implies that $|\mathcal{N}_{b,t,\alpha}| \leq 2|\mathcal{N}^+_{b,t,\alpha}|$. In the above, $\mathbf{E}$ stands for the standard expecation operator.

where the equality is due to the fact that $B^+(x^n, s^*)$ is a subset of $G(x^n)$. To this end, we observe from the definition of $B(x^n, s^*)$ that

$$\bigcup_{x^n \in \mathcal{N}^+_{b,t,\alpha}} B^+(x^n, s^*) \subseteq \bigcup_{r' \in \mathcal{T}_n(\mathcal{Y}): \|r' - s^*_{\mathcal{Y}}\|_1 \le \kappa/\sqrt{n}} T^n_{\mathcal{Y}}(r')$$

which in turn implies

$$\begin{aligned}\sum_{x^n \in \mathcal{N}^+_{b,t,\alpha}} |B^+(x^n, s^*)| &\le \sum_{\substack{r' \in \mathcal{T}_n(\mathcal{Y}): \\ \|r' - s^*_{\mathcal{Y}}\|_1 \le \kappa/\sqrt{n}}} |T^n_{\mathcal{Y}}(r')| \\ &\le 2^{nH(s^*_{\mathcal{Y}}) + \kappa_2\sqrt{n}} \end{aligned} \tag{6.14}$$

where κ_2 is a constant. In the above, the fact that for any $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$,

$$\begin{aligned}\|\pi - s^*\|_1 &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |\pi(x,y) - s^*(x,y)| \\ &\ge \sum_{y \in \mathcal{Y}} \left| \sum_{x \in \mathcal{X}} \pi(x,y) - \sum_{x \in \mathcal{X}} s^*(x,y) \right| \\ &= \|\pi_{\mathcal{Y}} - s^*_{\mathcal{Y}}\|_1 \end{aligned}$$

is used to upper bound $\bigcup_{x^n \in \mathcal{N}^+_{b,t,\alpha}} B^+(x^n, s^*)$. Putting (6.13) and (6.14) together, we get

$$\begin{aligned}&2^{nH(s^*_{\mathcal{Y}}) + \kappa_2\sqrt{n}} \\ &\quad \ge \sum_{x^n \in \mathcal{N}^+_{b,t,\alpha}} |B^+(x^n, s^*)| \\ &\quad \overset{1)}{\ge} |\mathcal{N}^+_{b,t,\alpha}| 2^{n[H(s^*) - H(t) + D(s^* \| t \circ P_{Y|X}) + \frac{\log \epsilon_n}{n}] - \kappa_0\sqrt{n}} \\ &\quad \overset{2)}{\ge} |\mathcal{N}_{b,t,\alpha}| 2^{n[H(s^*) - H(t) + D(s^* \| t \circ P_{Y|X}) + \frac{\log \epsilon_n}{n}] - \kappa_0\sqrt{n} - 1} \end{aligned} \tag{6.15}$$

where inequality 1) follows from (6.13), and inequality 2) is due to the inequality $|\mathcal{N}_{b,t,\alpha}| \le 2|\mathcal{N}^+_{b,t,\alpha}|$ derived above. It follows immediately from (6.15) that

$$\begin{aligned}|\mathcal{N}_{b,t,\alpha}| &\le 2^{n[H(t) + H(s^*_{\mathcal{Y}}) - H(s^*) - D(s^* \| t \circ P_{Y|X}) - \frac{\log \epsilon_n}{n}]} \\ &\quad \times 2^{(\kappa_0 + \kappa_2)\sqrt{n} + 1} \\ &\le 2^{n\left[H(t) - H_{\text{in}|\mathcal{Y}}(t \circ P_{Y|X}, \frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1}{n} | t)\right]} \\ &\quad \times 2^{\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \log \epsilon_n + (\kappa_0 + \kappa_2)\sqrt{n} + \kappa_1 + \kappa_1' + \kappa_1'' + 1}. \end{aligned} \tag{6.16}$$

Equation (6.16) gives the desired upper bound on the number of sequences (in $\Phi(\alpha)$ with type t) which the encoder can place into one bin without violating the decoding error probability constraint. From (6.16) and the fact that $|T^n_{\mathcal{X}}(t)| = 2^{nH(t) - \frac{|\mathcal{X}|-1}{2}\log n + O(1)}$, it follows that

$$\begin{aligned}&\frac{1}{n}\log\frac{p(t)}{p(b,1,t)} \\ &\quad = \frac{1}{n}\log\frac{|\{x^n : \tau(x^n) = t\}|}{|\mathcal{N}_{b,t,\alpha}|} \\ &\quad \ge H_{\text{in}|\mathcal{Y}}\left(t \circ P_{Y|X}, \frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1}{n} \middle| t\right) \\ &\qquad - \frac{1}{n}\log\left(\frac{2\epsilon_{b,t,\alpha} + \epsilon_n}{\epsilon_n}\right) - \frac{\kappa_0 + \kappa_2}{\sqrt{n}} \\ &\qquad - \left(\frac{|\mathcal{X}| - 1}{2}\right)\frac{\log n}{n} - O\left(\frac{1}{n}\right) \\ &\quad \overset{1)}{=} H(t \circ P_{Y|X}) - H((t \circ P_{Y|X})_{\mathcal{Y}}) \\ &\qquad + d_v(t)\sqrt{\frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1}{n}} \\ &\qquad - \frac{1}{n}\log\left(\frac{2\epsilon_{b,t,\alpha} + \epsilon_n}{\epsilon_n}\right) \\ &\qquad + O\left(\frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1}{n}\right) - \frac{\kappa_0 + \kappa_2}{\sqrt{n}} \\ &\qquad - \left(\frac{|\mathcal{X}| - 1}{2}\right)\frac{\log n}{n} - O\left(\frac{1}{n}\right) \\ &\quad \ge H(t \circ P_{Y|X}) - H((t \circ P_{Y|X})_{\mathcal{Y}}) \\ &\qquad + d_v(t)\sqrt{\frac{-\log(2\epsilon_{b,t,\alpha} + \epsilon_n) - \kappa_1}{n}} - O\left(\frac{-\log \epsilon_n}{n}\right) \\ &\qquad - \frac{\kappa_0 + \kappa_2}{\sqrt{n}} - \left(\frac{|\mathcal{X}| - 1}{2}\right)\frac{\log n}{n} - O\left(\frac{1}{n}\right) \end{aligned} \tag{6.17}$$

where the equality 1) follows from Lemma 4, and $d_v(t)$ is defined in (2.7).

To proceed, note that (6.17) holds uniformly for all $b \in \mathcal{I}$ and $t \in \Gamma_{\mathcal{X}}$. We average (6.17) with $p(b|1,t)$ over $b \in \mathcal{I}$. Observe that only one term on the right-hand-side of (6.17) depends on the index bin b. Note that $\sqrt{-\ln x - \xi}$ is a convex function when $0 < x < e^{-\frac{1}{2} - \xi}$ for $\xi > 0$. Thus if $\alpha < e^{-\frac{1}{2} - \kappa_1 \ln 2 - \ln 3}$, we can apply Jensen's inequality and get

$$\begin{aligned}&\frac{1}{n}\sum_{b \in \mathcal{I}} p(b|1,t)\log\frac{p(t)}{p(b,1,t)} \\ &\quad \ge H(t \circ P_{Y|X}) - H((t \circ P_{Y|X})_{\mathcal{Y}}) \\ &\qquad + d_v(t)\sqrt{\frac{-\log(2\epsilon_{t,\alpha} + \epsilon_n) - \kappa_1}{n}} + O\left(\frac{-\log \epsilon_n}{n}\right) \\ &\qquad - \frac{\kappa_0 + \kappa_2}{\sqrt{n}} - \left(\frac{|\mathcal{X}| - 1}{2}\right)\frac{\log n}{n} - O\left(\frac{1}{n}\right). \end{aligned} \tag{6.18}$$

We continue to average (6.18) with $p(1,t)$ over $t \in \mathcal{T}_n(\mathcal{X})$. (We can do this because $p(1,t) = 0$ when $t \notin \Gamma_{\mathcal{X}}$). Let us first calculate

$$\sum_{t \in \mathcal{T}_n(\mathcal{X})} p(1,t) d_v(t)\sqrt{\frac{-\log(2\epsilon_{t,\alpha} + \epsilon_n) - \kappa_1}{n}}.$$

Observe that although $\sqrt{\frac{-\log(2\epsilon_{t,\alpha} + \epsilon_n) - \kappa_1}{n}}$ is convex with respect to $\epsilon_{t,\alpha}$ when $\alpha < e^{-\frac{1}{2} - \kappa_1 \ln 2 - \ln 3}$, $d_v(t)$ is concave with respect to t. To address this problem, we note that

$$d_v(t) \ge d_v(P_X) - \kappa_5 \|t - P_X\|_1$$

for n sufficiently large, where $\kappa_5 > 0$ is a constant dependent only on P_{XY}. Thus

$$\begin{aligned}&\sum_{t \in \mathcal{T}_n(\mathcal{X})} p(1,t) d_v(t)\sqrt{\frac{-\log(2\epsilon_{t,\alpha} + \epsilon_n) - \kappa_1}{n}} \\ &\quad \ge \sum_{t \in \mathcal{T}_n(\mathcal{X})} p(1,t)[d_v(P_X) - \kappa_5 \|t - P_X\|_1] \end{aligned}$$

$$\times\sqrt{\frac{-\log(2\epsilon_{t,\alpha}+\epsilon_n)-\kappa_1}{n}}$$
$$\overset{1)}{\geq}\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(1,t)\left[d_v(P_X)\sqrt{\frac{-\log(2\epsilon_{t,\alpha}+\epsilon_n)-\kappa_1}{n}}\right.$$
$$\left.-\kappa_5\|t-P_X\|_1\sqrt{\frac{-\log\epsilon_n}{n}}\right]$$
$$\overset{2)}{\geq}\Pr\{A=1\}d_v(P_X)\sqrt{\frac{-\log(2\epsilon_\alpha+\epsilon_n)-\kappa_1}{n}}$$
$$-\kappa_5\sqrt{\frac{-\log\epsilon_n}{n}}\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(1,t)\|t-P_X\|_1$$
$$\geq\Pr\{A=1\}d_v(P_X)\sqrt{\frac{-\log(2\epsilon_\alpha+\epsilon_n)-\kappa_1}{n}}$$
$$-\kappa_5\sqrt{\frac{-\log\epsilon_n}{n}}\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)\|t-P_X\|_1. \tag{6.19}$$

In the above, inequality 1) is due to the fact that

$$\sqrt{\frac{-\log(2\epsilon_{t,\alpha}+\epsilon_n)-\kappa_1}{n}}<\sqrt{\frac{-\log\epsilon_n}{n}};$$

and inequality 2) follows from Jensen's inequality. At this moment, we invoke the following lemma, whose proof can be found in Appendix D.

Lemma 6: There exists a constant $\kappa_6>0$ such that $\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)\|t-P_X\|_1\leq\kappa_6/\sqrt{n}$ for all $n>1$.

It follows from (6.19) and Lemma 6 that

$$\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(1,t)d_v(t)\sqrt{\frac{-\log(2\epsilon_{t,\alpha}+\epsilon_n)-\kappa_1}{n}}$$
$$\geq\Pr\{A=1\}d_v(P_X)\sqrt{\frac{-\log(2\epsilon_\alpha+\epsilon_n)-\kappa_1}{n}}$$
$$-\kappa_7\frac{\sqrt{-\log\epsilon_n}}{n} \tag{6.20}$$

where $\kappa_7=\kappa_5\kappa_6$.

We next calculate

$$\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(1,t)[H(t\circ P_{Y|X})-H((t\circ P_{Y|X})_{\mathcal{Y}})]$$

Denote $[H(t\circ P_{Y|X})-H((t\circ P_{Y|X})_{\mathcal{Y}})]$ by $F(t)$. Expanding $F(t)$ at P_X by using Taylor's series, we have

$$F(t)=F(P_X)+\left.\frac{\partial F(t)}{\partial t}\right|_{t=P_X}(t-P_X)'$$
$$+\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'+o\left(\|t-P_X\|_1^2\right)$$
$$=H(X|Y)+\left.\frac{\partial F(t)}{\partial t}\right|_{t=P_X}(t-P_X)'$$
$$+\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'+o\left(\|t-P_X\|_1^2\right)$$

where t and P_X are regarded as row vectors, and $(t-P_X)'$ denotes the transpose of $(t-P_X)$. When $A=1$, we see that $\|t-P_X\|_1^2=O(\frac{\log n}{n})$. Thus

$$\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(1,t)[H(t\circ P_{Y|X})-H((t\circ P_{Y|X})_{\mathcal{Y}})]$$
$$\geq\sum_{t\in\Gamma_{\mathcal{X}}}p(t)[H(t\circ P_{Y|X})-H((t\circ P_{Y|X})_{\mathcal{Y}})]$$
$$-\Pr\{A=0,\tau(X^n)\in\Gamma_{\mathcal{X}}\}\log|\mathcal{X}|$$
$$\geq\sum_{t\in\Gamma_{\mathcal{X}}}p(t)\left[H(X|Y)+\left.\frac{\partial F(t)}{\partial t}\right|_{t=P_X}(t-P_X)'\right.$$
$$\left.+\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'\right]$$
$$-\Pr\{A=0,\tau(X^n)\in\Gamma_{\mathcal{X}}\}\log|\mathcal{X}|-o\left(\frac{\log n}{n}\right)$$
$$\geq H(X|Y)-\Pr\{A=0\}\log|\mathcal{X}|-o\left(\frac{\log n}{n}\right)$$
$$+\sum_{t\in\Gamma_{\mathcal{X}}}p(t)\left[\left.\frac{\partial F(t)}{\partial t}\right|_{t=P_X}(t-P_X)'\right.$$
$$\left.+\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'\right]$$
$$\overset{1)}{\geq}H(X|Y)-\Pr\{A=0\}\log|\mathcal{X}|-o\left(\frac{\log n}{n}\right)$$
$$+\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)\left[\left.\frac{\partial F(t)}{\partial t}\right|_{t=P_X}(t-P_X)'\right.$$
$$\left.+\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'\right]$$
$$\overset{2)}{=}H(X|Y)-\Pr\{A=0\}\log|\mathcal{X}|-o\left(\frac{\log n}{n}\right)$$
$$+\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'$$
$$\overset{3)}{\geq}H(X|Y)-\Pr\{A=0\}\log|\mathcal{X}|-o\left(\frac{\log n}{n}\right). \tag{6.21}$$

In the above, inequality 1) is due to $\Pr\{\tau(X^n)\notin\Gamma_{\mathcal{X}}\}\leq\frac{1}{n^2}$; and inequality 2) follows from that

$$\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)(t-P_X)=0.$$

In order to see that inequality 3) holds, we observe that for any $a_i,a_j\in\mathcal{X}$

$$\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)(t(a_i)-P_X(a_i))(t(a_j)-P_X(a_j))$$
$$=\begin{cases}\frac{1}{n}P_X(a_i)(1-P_X(a_i)), & \text{if } i=j\\ \frac{1}{n}P_X(a_i)P_X(a_j), & \text{if } i\neq j\end{cases}$$

and thus

$$\sum_{t\in\mathcal{T}_n(\mathcal{X})}p(t)\frac{1}{2}(t-P_X)\left.\frac{\partial^2F(t)}{\partial t^2}\right|_{t=P_X}(t-P_X)'=O\left(\frac{1}{n}\right).$$

In view of (6.18), (6.20), and (6.21), we see that

$$\begin{aligned}
&\frac{1}{n}\sum_{t\in\mathcal{T}_n(\mathcal{X})} p(1,t)\sum_{b\in\mathcal{I}} p(b|1,t)\log\frac{p(t)}{p(b,1,t)} \\
&\quad\ge H(X|Y)-\Pr\{A=0\}\log|\mathcal{X}|-o\left(\frac{\log n}{n}\right) \\
&\qquad+\Pr\{A=1\}d_v(P_X)\sqrt{\frac{-\log(2\epsilon_\alpha+\epsilon_n)-\kappa_1}{n}} \\
&\qquad-\kappa_7\frac{\sqrt{-\log\epsilon_n}}{n}-O\left(\frac{-\log\epsilon_n}{n}\right) \\
&\qquad-\frac{\kappa_0+\kappa_2}{\sqrt{n}}-\left(\frac{|\mathcal{X}|-1}{2}\right)\frac{\log n}{n}+O\left(\frac{1}{n}\right) \\
&\quad\ge H(X|Y)+d_v(P_X)\sqrt{\frac{-\log(2\epsilon_\alpha+\epsilon_n)-\kappa_1}{n}} \\
&\qquad-O\left(\frac{-\log\epsilon_n}{n}\right)-O\left(\frac{1}{\sqrt{n}}\right)-O(\epsilon_n). \qquad (6.22)
\end{aligned}$$

Since

$$\begin{aligned}
\epsilon_\alpha&=\frac{\epsilon_n-\Pr\{X^n\ne\hat{X}^n,A=0\}}{\Pr\{A=1\}} \\
&\le\frac{\epsilon_n}{\Pr\{A=1\}} \\
&\le\frac{\epsilon_n}{1-\frac{\epsilon_n}{\alpha}-\frac{1}{n^2}}
\end{aligned}$$

we have

$$\begin{aligned}
&\sqrt{\frac{-\log(2\epsilon_\alpha+\epsilon_n)-\kappa_1}{n}} \\
&\quad\ge\sqrt{\frac{-\log\epsilon_n-\log\left(\frac{2}{1-\frac{\epsilon_n}{\alpha}-\frac{1}{n^2}}+1\right)-\kappa_1}{n}} \\
&\quad\ge\sqrt{\frac{-\log\epsilon_n-\log 3+\log\left(1-\frac{\epsilon_n}{\alpha}-\frac{1}{n^2}\right)-\kappa_1}{n}} \\
&\quad\ge\sqrt{\frac{-\log\epsilon_n}{n}} \\
&\qquad+\frac{\log\left(1-\frac{\epsilon_n}{\alpha}-\frac{1}{n^2}\right)-\kappa_1-\log 3}{\sqrt{-n\log\epsilon_n}} \qquad (6.23)
\end{aligned}$$

where the last inequality follows from the observation that $\sqrt{x-y}\ge\sqrt{x}-y/\sqrt{x}$ for $0<y\le x$. Putting (6.23) back into (6.22), we finally get

$$\begin{aligned}
r(C_n)(X^n)&\ge H(X|Y)+d_v(P_X)\sqrt{\frac{-\log\epsilon_n}{n}} \\
&\quad-O\left(\frac{-\log\epsilon_n}{n}\right)-O\left(\frac{1}{\sqrt{n}}\right)-O(\epsilon_n) \\
&\quad-O\left(\frac{1}{\sqrt{-n\log\epsilon_n}}\right). \qquad (6.24)
\end{aligned}$$

Because of (6.24) and Lemma 4, this completes the proof of Theorem 1. □

Proof of Theorem 2: Theorem 2 can be proved in various ways. Here we only sketch a proof that is simplified from the proof of Theorem 1 and that of Lemma 5. Let $C_n=(f_n,g_n)$ be any fixed rate code as specified in Theorem 1. In view of the definition of the conditional intrinsic entropy in Section II and Lemma 2, we let s^* be a distribution in $\mathcal{P}^+(\mathcal{X}\times\mathcal{Y})$ such that

$$\begin{aligned}
H(s^*)-H(s^*_{\mathcal{Y}})&=H_{\text{in}|\mathcal{Y}}(P_{XY},\delta_n),\ \text{and} \\
D(s^*||P_{XY})&=\delta_n
\end{aligned}$$

where δ_n is a positive number to be specified later. Around s^*, we define a type set

$$\mathcal{T}_{s^*}\triangleq\left\{s\in\mathcal{T}_n(\mathcal{X}\times\mathcal{Y}):\|s-s^*\|_1\le\frac{1}{\sqrt{n}}\right\}.$$

Following an argument similar to that used to prove Lemma 5, we can show that

$$\begin{aligned}
&\Pr\{\tau(X^n,Y^n)\in\mathcal{T}_{s^*}\} \\
&\quad\ge 2^{-nD(s^*||P_{XY})-O(1)} \\
&\quad=2^{-n\delta_n-O(1)}. \qquad (6.25)
\end{aligned}$$

Select δ_n now so that the right-hand-side of (6.25) is equal to $2\epsilon_n$, where $\epsilon_n=P_e(C_n)$. That is

$$\delta_n=\frac{-\log\epsilon_n}{n}-O\left(\frac{1}{n}\right).$$

Since

$$\begin{aligned}
\epsilon_n&=\Pr\{\hat{X}^n\ne X^n\} \\
&\ge\Pr\{\hat{X}^n\ne X^n,\tau(X^n,Y^n)\in\mathcal{T}_{s^*}\}
\end{aligned}$$

we have

$$\Pr\{\hat{X}^n=X^n,\tau(X^n,Y^n)\in\mathcal{T}_{s^*}\}\ge\epsilon_n. \qquad (6.26)$$

For brevity, let us define

$$\begin{aligned}
B_{s^*}\triangleq|\{&(x^n,y^n)\in\mathcal{X}^n\times\mathcal{Y}^n: \\
&\tau(x^n,y^n)\in\mathcal{T}_{s^*},g_n(y^n,f_n(x^n))=x^n\}|.
\end{aligned}$$

It then follows from (6.26), the definition of $\mathcal{T}_{s^*}$, and an argument similar to that used in the proof of Lemma 5 that

$$\begin{aligned}
|B_{s^*}|&\ge\epsilon_n 2^{n[H(s^*)+D(s^*||P_{XY})]-\kappa_0\sqrt{n}} \\
&=2^{nH(s^*)-\kappa_0\sqrt{n}-O(1)} \qquad (6.27)
\end{aligned}$$

where κ_1 is a positive constant. Observe that there are at most $2^{nH(s^*_{\mathcal{Y}})+\kappa_2\sqrt{n}}$ side information sequences in the set B_{s^*}, where κ_2 is a positive constant. This implies that the code C_n needs at least

$$2^{n[H(s^*)-H(s^*_{\mathcal{Y}})]-(\kappa_0+\kappa_2)\sqrt{n}-O(1)} \qquad (6.28)$$

distinct codewords (or equivalently bins) to encode the source sequences in B_{s^*}. Equation (6.28), coupled with the fact that C_n is a fixed rate code and the definition of s^* above, further implies that

$$r(C_n)\ge H_{\text{in}|\mathcal{Y}}(P_{XY},\delta_n)-(\kappa_0+\kappa_2)\frac{1}{\sqrt{n}}-O\left(\frac{1}{n}\right).$$

Because of our selection of δ_n and Lemma 4, this completes the proof of Theorem 2. □

VII. Proof of Theorems 3 and 4

In this section, we prove Theorems 3 and 4.

Proof of Theorem 3: Let P_X and P_Y denote the marginals of P_{XY} over $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $P_{Y|X}$ denote the conditional probability distribution of Y given X. To prove this theorem, we shall construct a sequence of codes $\{C_n\}_{n=1}^{\infty}$ with desired decoding error probability and redundancy. Our construction directly reflects the knowledge that we learned from the analysis of the redundancy lower bound of SW coding in Section III, and makes conscious use of the marginal type of X^n.

To describe $C_n = (f_n, g_n)$, it suffices to see how f_n and g_n work. For any $t \in \mathcal{T}_n(\mathcal{X})$, and x^n with type t, the encoder f_n encodes x^n as follows.

Step 1: Encode t by using

$$\log|\mathcal{T}_n(\mathcal{X})| = \log\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} = O(\log n) \text{ bits.} \tag{7.29}$$

Step 2: If $\|t-P_X\|_1 \geq c_0\sqrt{\log n/n}$ where c_0 is a constant selected so that

$$\Pr\left\{\|\tau(X^n)-P_X\|_1 \geq c_0\sqrt{\frac{\log n}{n}}\right\} \leq \frac{1}{n^2} \tag{7.30}$$

encode x^n losslessly by using

$$\log|T^n_{\mathcal{X}}(t)| \leq [nH(t)] \text{ bits;} \tag{7.31}$$

otherwise, selects

$$R(t) = H_{\text{in}|\mathcal{Y}}(t\circ P_{Y|X}, \beta_n|t) + \alpha_n \tag{7.32}$$

where $\beta_n = O(\frac{-\log\epsilon_n}{n})$, and $\alpha_n = o(\sqrt{\frac{-\log\epsilon_n}{n}})$ are two positive numbers to be specified later. Construct $2^{nR(t)}$ bins. C_n then randomly selects a bin with probability $2^{-nR(t)}$, places x^n into the bin, and encodes the bin index by using $nR(t)$ bits. Note that the random bin selection is independent of (X,Y), and the codeword length $|f_n(x^n)|$ depends only on the type t of x^n.

On the decoder side, the decoding function g_n works as follows.

Step 1: Decode the type t from the transmitted codeword.

Step 2: If $\|t-P_X\|_1 \geq c_0\sqrt{\log n/n}$, decode the source sequence x^n from the transmitted codeword; otherwise, decode the bin index from the transmitted codeword, and continue to Step 3 below.

Step 3: This step is executed only if $\|t-P_X\|_1 < c_0\sqrt{\log n/n}$. For any side information sequence $y^n \in \mathcal{Y}^n$ and any $t \in \mathcal{T}_n(\mathcal{X})$, define $T_{y^n,t}$ by

$$T_{y^n,t} \triangleq \left\{x^n \in T^n_{\mathcal{X}}(t) : D(\tau(x^n,y^n)\|t\circ P_{Y|X}) \leq \beta_n\right\}.$$

Note that $T_{y^n,t}$ could be empty. Our code C_n then reconstructs $\hat{x}^n$ as a sequence in the bin specified by the transmitted bin index that is also in $T_{y^n,t}$.[5] That is, if we use BIN_i to denote the set of all sequences from $\mathcal{X}$ in the ith bin, then

$$\hat{x}^n \in T_{y^n,t} \bigcap \text{BIN}_{i^*}$$

if $T_{y^n,t}\bigcap \text{BIN}_{i^*}$ is non-empty, where i^* denotes the bin index sent from the encoder; otherwise, $\hat{x}^n$ is selected arbitrarily. If the set $T_{y^n,t}\bigcap \text{BIN}_{i^*}$ consists of more than one sequence, C_n selects an arbitrary one in $T_{y^n,t}\bigcap \text{BIN}_{i^*}$ as $\hat{x}^n$.

Suppose that $\tau(X^n) = t$, where $\|t-P_X\|_1 < c_0\sqrt{\log n/n}$. In view of the encoding and decoding process of C_n described above, we see that a decoding error happens if either one of the following two events occurs.

i) $X^n \notin T_{Y^n,t}$.

ii) $X^n \in T_{Y^n,t}$, but there exist other sequences in $T_{Y^n,t}$ that also fall in the bin containing X^n.

In view of these, we have for $t \in \{t \in \mathcal{T}_n(\mathcal{X}) : \|t-P_X\|_1 < c_0\sqrt{\log n/n}\}$

$$\begin{aligned}\Pr\{\hat{X}^n \neq X^n|\tau(X^n)=t\} &= \Pr\{X^n \notin T_{Y^n,t}|\tau(X^n)=t\} \\ &\quad + \Pr\{X^n \in T_{Y^n,t}|\tau(X^n)=t\} \\ &\quad \times \Pr\{\hat{X}^n \neq X^n|X^n \in T_{Y^n,t}\}.\end{aligned} \tag{7.33}$$

In the following, our job is to upper bound the two terms on the right hand side of (7.33). The derivation of these bounds makes use of the concept of intrinsic conditional entropy.

To upper bound $\Pr\{X^n \notin T_{Y^n,t}|\tau(X^n)=t\}$, we define

$$\mathcal{S}(t,\beta_n) \triangleq \{s \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y}) : D(s\|t\circ P_{Y|X}) > \beta_n \text{ and } s_{\mathcal{X}} = t\}.$$

Let x^n denote a sequence in $T^n_{\mathcal{X}}(t)$. Then

$$\begin{aligned}&\Pr\{X^n \notin T_{Y^n,t}|X^n=x^n\} \\ &= \sum_{y^n\in\mathcal{Y}^n:\, x^n\notin T_{y^n,t}} \Pr\{Y^n=y^n|X^n=x^n\} \\ &= \sum_{s\in\mathcal{S}(t,\beta_n)}\ \sum_{y^n\in\mathcal{Y}^n:\, \tau(x^n,y^n)=s} \Pr\{Y^n=y^n|X^n=x^n\} \\ &= \sum_{s\in\mathcal{S}(t,\beta_n)} |\{y^n\in\mathcal{Y}^n :\, \tau(x^n,y^n)=s\}| \\ &\quad \times 2^{-n[H(s)-H(t)+D(s\|t\circ P_{Y|X})]} \\ &= \sum_{s\in\mathcal{S}(t,\beta_n)} 2^{-nD(s\|t\circ P_{Y|X})-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log n+O(1)}\end{aligned} \tag{7.34}$$

[5]The decoding rule described in Step 3 of the decoding function g_n is similar to the standard joint typicality decoding rule with two subtle differences: 1) the joint type must satisfy the marginal constraint, i.e., the set $T_{y^n,t}$ consists of only sequences with the same type known to the decoder; 2) and the typicality is evaluated by using relative entropy instead of L1 norm. The main reason for these subtle touches on the decoding rule is to derive a sharp upper bound of the redundancy of SW coding, while at the same time maintains the simplicity compared to the *maximum a posteriori*(MAP) decoding rule.

In the above, the last equality follows from Stirling's approximation of integer factorials (see (C3) in Appendix C for details).

To continue, we partition $\mathcal{S}(t,\beta_n)$ into a series of mutually exclusive subsets $\{\mathcal{S}_k(t,\beta_n)\}_{k=0}^{k_{\max}}$ as follows:

$$\mathcal{S}_k(t,\beta_n) \triangleq \{s \in \mathcal{S}(t,\beta_n) : k \le n[D(s\|t\circ P_{Y|X}) - \beta_n] < k+1\}$$

where $k_{\max} \triangleq \min\{k : \bigcup_{i=0}^{k} \mathcal{S}_i(t,\beta) = \mathcal{S}(t,\beta_n)\}$. Thus the sum $\sum_{s\in\mathcal{S}(t,\beta_n)} 2^{-nD(s\|t\circ P_{Y|X})}$ on the right-hand side of (7.34) can be upper-bounded by

$$\begin{aligned}
&\sum_{s\in\mathcal{S}(t,\beta_n)} 2^{-nD(s\|t\circ P_{Y|X})} \\
&\quad= \sum_{k=0}^{k_{\max}} \sum_{s\in\mathcal{S}_k(t,\beta_n)} 2^{-nD(s\|t\circ P_{Y|X})} \\
&\quad\le \sum_{k=0}^{k_{\max}} \sum_{s\in\mathcal{S}_k(t,\beta_n)} 2^{-n\beta_n-k} \\
&\quad= 2^{-n\beta_n} \sum_{k=0}^{k_{\max}} |\mathcal{S}_k(t,\beta_n)| 2^{-k}. \qquad (7.35)
\end{aligned}$$

It remains to upper bound $|\mathcal{S}_k(t,\beta_n)|$, $0 \le k \le k_{\max}$. Examining the definition of $\mathcal{S}_k(t,\beta_n)$ above, we find that for $s \in \mathcal{S}_k(t,\beta_n)$

$$\begin{aligned}
\frac{1}{2\ln 2}\|s - t\circ P_{Y|X}\|_1^2 &\le D(s\|t\circ P_{Y|X}) \\
&\le \frac{k+1}{n} + \beta_n \qquad (7.36)
\end{aligned}$$

where the first inequality follows from the L1 bound on relative entropy [3, Lemma 12.6.1]. This bound implies that

$$\begin{aligned}
|\mathcal{S}_k(t,\beta_n)| &\le 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(n(k+1)+n^2\beta_n))+O(1)} \\
&= 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log n+\log(k+1+n\beta_n))+O(1)}. \qquad (7.37)
\end{aligned}$$

Combining (7.34), (7.35), and (7.37), we arrive at

$$\begin{aligned}
&\Pr\{X^n \notin T_{Y^n,t} | X^n = x^n\} \\
&\le 2^{-n\beta_n} \sum_{k=0}^{k_{\max}} 2^{-k+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(k+1+n\beta_n))+O(1)} \\
&= 2^{-n\beta_n+O(1)} \\
&\quad\times \sum_{k=0}^{k_{\max}} 2^{-k+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log(k+1)+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\left(\log(1+\frac{n\beta_n}{k+1})\right)} \\
&\le 2^{-n\beta_n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(1+n\beta_n))+O(1)} \\
&\quad\times \sum_{k=0}^{k_{\max}} 2^{-k+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log(k+1)} \\
&\le 2^{-n\beta_n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(1+n\beta_n))+O(1)} \\
&\quad\times \int_0^\infty 2^{-x}(1+x)^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}} dx \\
&\le 2^{-n\beta_n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(1+n\beta_n))+O(1)} \\
&\quad\times \int_0^\infty 2^{-x}(1+x)^{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|} dx \\
&= 2^{-n\beta_n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(1+n\beta_n))+O(1)} \\
&\quad\times \sum_{m=0}^{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|} \left[\frac{1}{(m!)(\ln 2)^{m+1}}\right] \\
&= 2^{-n\beta_n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(1+n\beta_n))+O(1)} \qquad (7.38)
\end{aligned}$$

where $m!$ denotes the integer factorial with convention $0! = 1$, and the last equality follows from our assumption of finite alphabets.

Leaving the selection of β_n in (7.38) to a later moment, we turn our attention to $\Pr\{\hat{X}^n \ne X^n | X^n \in T_{Y^n,t}\}$. Suppose that $Y^n = y^n$. Given $X^n \in T_{y^n,t}$, the probability of the event $\hat{X}^n \ne X^n$ is upper bounded by the probability that another sequence z^n from $T_{y^n,t}$ is placed in the same bin as X^n. Recall that each bin is selected with uniform probability $2^{-nR(t)}$, and that the bin selection is independent of (X,Y). Define

$$\begin{aligned}
\mathcal{S}^+(t,\beta_n) \triangleq \{&s \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y}) : \\
&D(s\|t\circ P_{Y|X}) \le \beta_n \text{ and } s_{\mathcal{X}} = t\}
\end{aligned}$$

and

$$M^* \triangleq \max_{y^n\in\mathcal{Y}^n:\, \tau(x^n,y^n)\in\mathcal{S}^+(t,\beta_n)} |T_{y^n,t}|.$$

Then a standard argument can be used to show that

$$\Pr\{\hat{X}^n \ne x^n | X^n \in T_{Y^n,t}, X^n = x^n\} \le M^* 2^{-nR(t)}. \qquad (7.39)$$

It is clear that in order to upper bound (7.39), it suffices to upper bound M^*. We hint at this moment that our bound on M^* is related to intrinsic conditional entropy. Toward this direction, we see that from an argument similar to that used in the proof of Theorem 1, it follows that for any $s \in \mathcal{S}^+(t,\beta_n)$ and $y^n \in T_{\mathcal{Y}}^n(s_{\mathcal{Y}})$,

$$\begin{aligned}
&|\{x^n \in T_{\mathcal{X}}^n(t) : \tau(x^n,y^n) = s\}| \\
&\qquad= 2^{n[H(s)-H(s_{\mathcal{Y}})]-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{Y}|}{2}\log n+O(1)}. \qquad (7.40)
\end{aligned}$$

Let s^* denote the type in $\mathcal{S}^+(t,\beta_n)$ such that

$$s^* = \operatorname*{argmax}_{s\in\mathcal{S}^+(t,\beta_n)} [H(s) - H(s_{\mathcal{Y}})].$$

Then for each y^n with nonempty $T_{y^n,t}$, we have

$$\begin{aligned}
&|T_{y^n,t}| \\
&\quad= \sum_{s\in\mathcal{S}^+(t,\beta_n):s_{\mathcal{Y}}=\tau(y^n)} |\{x^n \in T_{\mathcal{X}}^n(t) :\ \tau(x^n,y^n) = s\}| \\
&\quad\overset{1)}{=} \sum_{\substack{s\in\mathcal{S}^+(t,\beta_n):\\ s_{\mathcal{Y}}=\tau(y^n)}} 2^{n[H(s)-H(s_{\mathcal{Y}})]-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{Y}|}{2}\log n+O(1)} \\
&\quad\le |\{s \in \mathcal{S}^+(t,\beta_n) : s_{\mathcal{Y}} = \tau(y^n)\}| \\
&\qquad\times 2^{n[H(s^*)-H(s^*_{\mathcal{Y}})]-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{Y}|}{2}\log n+O(1)}. \qquad (7.41)
\end{aligned}$$

In the above, the equality 1) is due to (7.40). Using the argument that led to (7.37), we can upper-bound

$$|\{s \in \mathcal{S}^+(t,\beta_n) : s_{\mathcal{Y}} = \tau(y^n)\}|$$

by

$$\left|\{s \in \mathcal{S}^+(t,\beta_n) : s_{\mathcal{Y}} = \tau(y^n)\}\right| \le 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}(\log n+\log n\beta_n)+O(1)}. \quad (7.42)$$

Putting (7.42) back into (7.41) leads to

$$\begin{aligned}|T_{y^n,t}| &\le 2^{n[H(s^*)-H(s^*_{\mathcal{Y}})]-\frac{|\mathcal{X}|-1}{2}\log n} \times 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}\log n\beta_n+O(1)} \\ &\le 2^{nH_{\text{in}|\mathcal{Y}}(t\circ P_{Y|X},\beta_n|t)-\frac{|\mathcal{X}|-1}{2}\log n} \times 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}\log n\beta_n+O(1)} \quad (7.43)\end{aligned}$$

where the last inequality follows from the definition of intrinsic conditional entropy in (2.3), and the definitions of s^* and $\mathcal{S}^+(t,\beta_n)$ above. Since (7.43) holds for any y^n and x^n such that $\tau(x^n,y^n) \in \mathcal{S}^+(t,\beta_n)$, it follows that

$$M^* \le 2^{nH_{\text{in}|\mathcal{Y}}(t\circ P_{Y|X},\beta_n|t)-\frac{|\mathcal{X}|-1}{2}\log n} \times 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}\log n\beta_n+O(1)} \quad (7.44)$$

which together with (7.39), implies

$$\begin{aligned}&\Pr\{\hat{X}^n \ne x^n | X^n \in T_{Y^n,t}, X^n = x^n\} \\ &\quad \le 2^{n[H_{\text{in}|\mathcal{Y}}(t\circ P_{Y|X},\beta_n|t)-R(t)]} \times 2^{-\frac{|\mathcal{X}|-1}{2}\log n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}\log n\beta_n+O(1)} \\ &\quad \le 2^{-n\alpha_n-\frac{|\mathcal{X}|-1}{2}\log n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}\log n\beta_n+O(1)} \quad (7.45)\end{aligned}$$

where the last inequality is due to (7.32).

For convenience, we now combine (7.33), (7.38), and (7.45) into

$$\begin{aligned}&\Pr\{X^n \ne \hat{X}^n | \tau(X^n) = t\} \\ &\quad \le 2^{-n\beta_n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}(\log(1+n\beta_n))+O(1)} \\ &\qquad + 2^{-n\alpha_n-\frac{|\mathcal{X}|-1}{2}\log n+\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}\log n\beta_n+O(1)}. \quad (7.46)\end{aligned}$$

In view of (7.46), we make the selections of β_n and α_n as follows:

$$\begin{cases}\beta_n = \frac{-\log\epsilon_n}{n} + \kappa_1\left(\frac{\log(-\log\epsilon_n)}{n}\right) \\ \alpha_n = \frac{-\log\epsilon_n}{n} + \kappa_2\left(\frac{\log(-\log\epsilon_n)}{n}\right)\end{cases} \quad (7.47)$$

where $\kappa_1 > \frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}$ and $\kappa_2 > \frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|-|\mathcal{Y}|+1}{2}$ are constants selected so that the two terms on the right-hand-side of (7.46) are upper bounded by $\frac{\epsilon_n}{2}$, respectively. Consequently, it follows from (7.46) and the description of our code C_n above that

$$\begin{aligned}&\Pr\{X^n \ne \hat{X}^n\} \\ &\quad = \sum_{t\in\mathcal{T}_n(\mathcal{X})} \Pr\{\tau(X^n)=t\}\Pr\{X^n \ne \hat{X}^n | \tau(X^n)=t\} \\ &\quad \le \sum_{t\in\mathcal{T}_n(\mathcal{X}):\|t-P_X\|_1\le c_0\log n/n} \Pr\{\tau(X^n)=t\}\epsilon_n \\ &\quad \le \epsilon_n \quad (7.48)\end{aligned}$$

for sufficiently large n.

To finish the proof, we need to upper bound the average compression rate of the code C_n. Note that even though C_n is a random code by construction, the codeword length $|f_n(x^n)|$, for each x^n, depends only on the type t of x^n. Thus, for any realization of C_n, we have

$$\begin{aligned}r(C_n) &\le O\left(\frac{\log n}{n}\right) \\ &\quad + \sum_{t\in\mathcal{T}_n(\mathcal{X}):\|t-P_X\|_1<c_0\log n/n} \Pr\{\tau(X^n)=t\}R(t) \\ &\quad + \sum_{t\in\mathcal{T}_n(\mathcal{X}):\|t-P_X\|_1\ge c_0\log n/n} \Pr\{\tau(X^n)=t\}H(t) \\ &\le O\left(\frac{\log n}{n}\right) + \frac{\log|\mathcal{X}|}{n^2} \\ &\quad + \sum_{t\in\mathcal{T}_n(\mathcal{X}):\|t-P_X\|_1<c_0\sqrt{\log n/n}} \Pr\{\tau(X^n)=t\} \\ &\quad \times [H_{\text{in}|\mathcal{Y}}(t\circ P_{Y|X},\beta_n|t)+\alpha_n] \\ &\overset{1)}{\le} O\left(\frac{\log n}{n}\right) + H_{\text{in}|\mathcal{Y}}(P_{XY},\beta_n|P_X)+\alpha_n \\ &\overset{2)}{=} O\left(\frac{\log n}{n}\right) + H(X|Y) + d_v\sqrt{\beta_n} + O(\beta_n) + \alpha_n \\ &\overset{3)}{=} H(X|Y) + d_v\sqrt{\frac{-\log\epsilon_n}{n}} \\ &\quad + O\left(\left(\frac{\log(-\log\epsilon_n)}{-\log\epsilon_n}\right)\sqrt{\frac{-\log\epsilon_n}{n}}\right) \\ &\quad + O\left(\frac{-\log\epsilon_n}{n}\right) + O\left(\frac{\log(-\log\epsilon_n)}{n}\right). \quad (7.49)\end{aligned}$$

In the above, the inequality 1) is due to the concavity of $H_{\text{in}|\mathcal{Y}}(\pi,\delta|\pi_{\mathcal{X}})$ with respect to π (see Lemma 2); the equality 2) follows from Lemma 4 and its proof in Appendix B; and the equality 3) is obtained by plugging in our selections of β_n and α_n in (7.47).

Combining (7.48) with (7.49) now implies that there exists an order n deterministic variable rate SW code with the decoding error probability less than or equal to ϵ_n and the average compression rate upper bounded by (7.49). This completes the proof of Theorem 3. □

Proof of Theorem 4: In order to prove Theorem 4, we modify the code constructed in the proof of Theorem 3 so that it becomes a fixed rate code as follows. Specifically, in Step 2 of the encoder f_n, for all type $t \in \mathcal{T}_n(\mathcal{X})$, select

$$R = H_{\text{in}|\mathcal{Y}}(P_{XY},\beta_n) + \alpha_n$$

where β_n and α_n are selected in (7.47). Our modified code C_n then randomly selects a bin with probability 2^{-nR}, places x^n into the bin, and encodes the bin index by using nR bits. Since in Step 1, the type t is encoded into a constant number of bits, the modified code C_n is indeed a fixed rate code. Using an argument similar to that used in the proof of Theorem 3, we can easily show that for sufficiently large n

$$\Pr\{\hat{X}^n \ne X^n\} \le \epsilon_n$$

with the modified code C_n. This, coupled with the definition of R and Lemma 4, completes the proof of Theorem 4. □

VIII. Conclusion

In this paper, we have characterized the compression performance in bits per symbol of both variable rate and fixed rate SW coding for memoryless source-side information pairs up to the second order inclusive when the decoding error probability ϵ_n goes to 0 fast enough, but not exponentially as the block length $n \to \infty$. The characterization, on one hand, implies that surprisingly, variable rate SW coding and fixed rate SW coding approach asymptotically the same compression limit at a speed of the same order $O(\sqrt{\frac{-\log \epsilon_n}{n}})$. This sharply contrasts with the fact that in classical lossless coding, the redundancy of zero-error variable rate coding, which is in the order $O(1/n)$, is much better than that of fixed rate coding, which is in the order $\Theta(\sqrt{\frac{-\log \epsilon_n}{n}})$. On the other hand, the characterization also implies that for large finite block lengths n, variable rate SW coding is indeed more efficient than fixed rate SW coding in general. A necessary and sufficient condition has also been derived under which variable rate SW coding and fixed rate SW coding have the same compression performance up to the second order inclusive. The design of practical SW codes with finite block lengths is not simply a matter of approaching the conditional entropy rate $H(X|Y)$; instead, it is more about the tradeoff among the compression rate, decoding error probability, and block length. During the course of proving our main results, new information quantities called intrinsic entropy and intrinsic conditional entropy have been introduced and analyzed. It is expected that these information quantities will have applications to other problems in information theory as well such as SW coding with multiple encoders.

Appendix A

In this section, we prove Lemma 3. For any $0 \le \lambda \le 1$, define s_λ by

$$s_\lambda = (1-\lambda)P_{XY} + \lambda P_X \circ P_Y.$$

Note that the marginals of s_λ over $\mathcal{X}$ and $\mathcal{Y}$ are P_X and P_Y, respectively. Then on the one hand we have

$$\begin{aligned} H(s_\lambda) - H(P_{XY}) &\ge \lambda(H(P_X) + H(P_Y) - H(P_{XY})) \\ &\ge \lambda\kappa \end{aligned} \tag{A1}$$

where κ is a positive constant. In the above, the last inequality follows from the assumption that $I(X;Y) > 0$. On the other hand, when λ is small, there exists a constant κ_1 such that

$$\begin{aligned} D(s_\lambda \| P_{XY}) &\le \kappa_1 \lambda^2 \| P_{XY} - P_X \circ P_Y \|_1^2 \\ &\le \kappa_1 |\mathcal{X}|^2 |\mathcal{Y}|^2 \lambda^2. \end{aligned} \tag{A2}$$

Select $\lambda \sim \frac{1}{|\mathcal{X}||\mathcal{Y}|}\sqrt{\frac{\delta}{\kappa_1}}$. Because of (A1), (A2), and the fact that $H_{\text{in}}(P_{XY}, \delta) \ge H(s_\lambda)$ by definition, this completes the proof of Lemma 3. □

Appendix B

In this section, we prove Lemma 4. For brevity, let $\pi = P_{XY}$. In view of (2.2) and Lemma 2, we see that in order to prove the first part of Lemma 4, it suffices to solve the following constrained maximization problem

$$\max H(s) - H(s_{\mathcal{Y}}) - H(\pi) + H(\pi_{\mathcal{Y}}) \quad \text{subject to } D(s\|\pi) = \delta. \tag{B1}$$

Let $\delta_s(x,y) \triangleq s(x,y) - \pi(x,y)$. From [3, Lemma 12.6.1], we see that

$$\|s - \pi\|_1^2 \le 2(\ln 2)\delta.$$

Thus, when δ is small, we can use Taylor's series to expand $H(s) - H(s_{\mathcal{Y}})$ at π, and get

$$\begin{aligned} &(\ln 2)[H(s) - H(s_{\mathcal{Y}})] \\ &\quad = H(\pi)\ln 2 - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \delta_s(x,y)\ln\pi(x,y) - H(\pi_{\mathcal{Y}})\ln 2 \\ &\qquad + \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}}\delta_s(x,y)\right]\ln\pi_{\mathcal{Y}}(y) + O(\delta). \end{aligned} \tag{B2}$$

Similarly, we expand $D(s\|\pi)$ at π, and get

$$D(s\|\pi) = \frac{1}{2\ln 2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\delta_s^2(x,y)}{\pi(x,y)} + O(\delta^{3/2}). \tag{B3}$$

In view of (B2), and (B3), we see that when δ is small, we can simplify (B1) into the following constrained maximization problem.

$$\begin{aligned} \max \quad & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\delta_s(x,y)\ln\pi(x,y) \\ & \qquad + \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}}\delta_s(x,y)\right]\ln\pi_{\mathcal{Y}}(y) \\ \text{subject to} \quad & \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\delta_s^2(x,y)}{\pi(x,y)} = (2\ln 2)\delta \\ & \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\delta_s(x,y) = 0 \end{aligned}$$

To solve the above problem, we can use the standard Lagrange multiplier. Define

$$\begin{aligned} J_s \triangleq & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\delta_s(x,y)\ln\pi(x,y) \\ & + \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}}\delta_s(x,y)\right]\ln\pi_{\mathcal{Y}}(y) \\ & - \lambda\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\delta_s^2(x,y)}{\pi(x,y)} \\ & - \alpha\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\delta_s(x,y). \end{aligned} \tag{B4}$$

The first derivative of J_s with respect to $\delta_s(x,y)$ is given by

$$\frac{\partial J_s}{\partial \delta_s(x,y)} = -\ln \pi(x,y) + \ln \pi_{\mathcal{Y}}(y) - 2\lambda \frac{\delta_s(x,y)}{\pi(x,y)} - \alpha. \quad \text{(B5)}$$

Letting (B5) equal zero, we have

$$\begin{aligned}\delta_s(x,y) &= \frac{\pi(x,y)}{2\lambda}[-\ln \pi(x,y) + \ln \pi_{\mathcal{Y}}(y) - \alpha] \\ &= \frac{\pi(x,y)}{2\lambda}\left[\ln \frac{\pi_{\mathcal{Y}}(y)}{\pi(x,y)} - \alpha\right]. \end{aligned} \quad \text{(B6)}$$

(B6), together with the constraints in (B4), leads to

$$\alpha = H(X|Y)\ln 2, \quad \text{(B7)}$$

and

$$2\lambda = \sqrt{\frac{1}{(2\ln 2)\delta}\left[\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \pi(x,y)\ln^2 \frac{\pi_{\mathcal{Y}}(y)}{\pi(x,y)} - \alpha^2\right]}. \quad \text{(B8)}$$

Note that the difference within the brackets above is positive whenever $I(X;Y) > 0$ or X is not uniformly distributed over $\mathcal{X}$. Consequently

$$\begin{aligned}&-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \delta_s(x,y)\ln \pi(x,y) \\ &\quad + \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} \delta_s(x,y)\right]\ln \pi_{\mathcal{Y}}(y) \\ &\quad = -\frac{1}{2\lambda}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \pi(x,y) \\ &\qquad \times [-\ln \pi(x,y) + \ln \pi_{\mathcal{Y}}(y) - \alpha]\ln \pi(x,y) \\ &\qquad + \frac{1}{2\lambda}\sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} \pi(x,y)\right. \\ &\qquad \left.\times[-\ln \pi(x,y) + \ln \pi_{\mathcal{Y}}(y) - \alpha]\right]\ln \pi_{\mathcal{Y}}(y) \\ &\quad = \frac{1}{2\lambda}\left[\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \pi(x,y)\ln^2 \frac{\pi_{\mathcal{Y}}(y)}{\pi(x,y)} - \alpha^2\right] \\ &\quad = \sqrt{(2\ln 2)\delta} \\ &\qquad \times \sqrt{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \pi(x,y)\ln^2 \frac{\pi_{\mathcal{Y}}(y)}{\pi(x,y)} - H^2(X|Y)(\ln 2)^2}. \end{aligned} \quad \text{(B9)}$$

This completes the proof of the first part of Lemma 4.

We now use the same strategy to prove the second part of Lemma 4. For brevity, let $\pi = t \circ P_{Y|X}$ from this moment on. In view of (2.3) and Lemma 2, we see that in order to prove Lemma 4, it suffices to solve the following constrained maximization problem:

$$\max H(s) - H(s_{\mathcal{Y}}) - H(\pi) + H(\pi_{\mathcal{Y}}) \quad \text{subject to } D(s\|\pi) = \delta,\ s_{\mathcal{X}} = t. \quad \text{(B10)}$$

Observe that

$$D(s\|\pi) = \sum_{x\in\mathcal{X}} t(x) D(s(\cdot|x)\|\pi(\cdot|x))$$

where $s(y|x) = s(x,y)/t(x)$ and $\pi(y|x) = \pi(x,y)/t(x) = P_{Y|X}(y|x)$ for any $y \in \mathcal{Y}$. Let $\delta_s(y|x) \triangleq s(y|x) - \pi(y|x)$. From [3, Lemma 12.6.1], we see that

$$\sum_{x\in\mathcal{X}} t(x)\left[\sum_{y\in\mathcal{Y}} |\delta_s(y|x)|\right]^2 \le 2(\ln 2)\delta$$

where $\ln$ denotes the natural logarithm. Thus, when δ is small, we can use Taylor's series to expand $H(s) - H(s_{\mathcal{Y}})$ at π, and get

$$\begin{aligned}&(\ln 2)[H(s) - H(s_{\mathcal{Y}})] \\ &\quad = H(\pi)\ln 2 - \sum_{x\in\mathcal{X}} t(x)\sum_{y\in\mathcal{Y}} \delta_s(y|x)\ln \pi(y|x) - H(\pi_{\mathcal{Y}})\ln 2 \\ &\qquad + \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} t(x)\delta_s(y|x)\right]\ln \pi_{\mathcal{Y}}(y) + O(\delta). \end{aligned} \quad \text{(B11)}$$

Similarly, we expand $D(s\|\pi)$ at π, and get

$$D(s\|\pi) = \frac{1}{2\ln 2}\sum_{x\in\mathcal{X}} t(x)\sum_{y\in\mathcal{Y}} \frac{\delta_s^2(y|x)}{\pi(y|x)} + O(\delta^{3/2}). \quad \text{(B12)}$$

In view of (B11) and (B12), we see that when δ is small, we can simplify (B10) into the following constrained maximization problem:

$$\begin{aligned}\max \quad & -\sum_{x\in\mathcal{X}} t(x)\sum_{y\in\mathcal{Y}} \delta_s(y|x)\ln \pi(y|x) \\ &\qquad + \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} t(x)\delta_s(y|x)\right]\ln \pi_{\mathcal{Y}}(y) \\ \text{subject to} \quad & \sum_{x\in\mathcal{X}} t(x)\sum_{y\in\mathcal{Y}} \frac{\delta_s^2(y|x)}{\pi(y|x)} = (2\ln 2)\delta \\ &\qquad \sum_{y\in\mathcal{Y}} \delta_s(y|x) = 0 \text{ for any } x \in \mathcal{X}. \end{aligned} \quad \text{(B13)}$$

To solve the above problem, we again use the standard Lagrange multiplier. Define

$$\begin{aligned}J_s \triangleq & -\sum_{x\in\mathcal{X}} t(x)\sum_{y\in\mathcal{Y}} \delta_s(y|x)\ln \pi(y|x) \\ &+ \sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} t(x)\delta_s(y|x)\right]\ln \pi_{\mathcal{Y}}(y) \\ &- \lambda\sum_{x\in\mathcal{X}} t(x)\sum_{y\in\mathcal{Y}} \frac{\delta_s^2(y|x)}{\pi(y|x)} \\ &- \sum_{x\in\mathcal{X}} \alpha_x \sum_{y\in\mathcal{Y}} \delta_s(y|x). \end{aligned} \quad \text{(B14)}$$

The first derivative of J_s with respect to $\delta_s(y|x)$ is given by

$$\begin{aligned}\frac{\partial J_s}{\partial \delta_s(y|x)} = & -t(x)\ln \pi(y|x) \\ &+ t(x)\ln \pi_{\mathcal{Y}}(y) - 2\lambda \frac{t(x)\delta_s(y|x)}{\pi(y|x)} - \alpha_x. \end{aligned} \quad \text{(B15)}$$

Letting (B15) equal zero, we have

$$\delta_s(y|x) = \frac{\pi(y|x)}{2\lambda}\left[-\ln \pi(y|x) + \ln \pi_{\mathcal{Y}}(y) - \frac{\alpha_x}{t(x)}\right]. \tag{B16}$$

(B16), together with the constraints in (B13), leads to (B17) and (B18) shown at the bottom of the page. Note that the difference within the brackets above is positive whenenver $I(X;Y) > 0$. Consequently

$$\begin{aligned}
&-\sum_{x\in\mathcal{X}} t(x) \sum_{y\in\mathcal{Y}} \delta_s(y|x) \ln \pi(y|x) \\
&\quad + \sum_{y\in\mathcal{Y}} \left[\sum_{x\in\mathcal{X}} t(x)\delta_s(y|x)\right] \ln \pi_{\mathcal{Y}}(y) \\
&\qquad = \frac{-1}{2\lambda} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \pi(x,y)[-\ln \pi(y|x) + \ln \pi_{\mathcal{Y}}(y) \\
&\qquad\quad + (\ln 2) D(\pi(\cdot|x)||\pi_{\mathcal{Y}})] \ln \pi(y|x) \\
&\qquad\quad + \frac{1}{2\lambda} \sum_{y\in\mathcal{Y}} \left[\sum_{x\in\mathcal{X}} \pi(x,y)[-\ln \pi(y|x) + \ln \pi_{\mathcal{Y}}(y) \right. \\
&\qquad\quad \left. + (\ln 2) D(\pi(\cdot|x)||\pi_{\mathcal{Y}})]\right] \ln \pi_{\mathcal{Y}}(y) \\
&\qquad = \frac{-1}{2\lambda} \left\{ \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \pi(x,y)[-\ln^2 \pi(y|x) \right. \\
&\qquad\quad + \ln \pi_{\mathcal{Y}}(y) \ln \pi(y|x) \\
&\qquad\quad + (\ln 2) D(\pi(\cdot|x)||\pi_{\mathcal{Y}}) \ln \pi(y|x)] \\
&\qquad\quad - \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \pi(x,y)[-\ln \pi(y|x) \\
&\qquad\quad \times \ln \pi_{\mathcal{Y}}(y) + \ln^2 \pi_{\mathcal{Y}}(y) \\
&\qquad\quad \left. + (\ln 2) D(\pi(\cdot|x)||\pi_{\mathcal{Y}}) \ln \pi_{\mathcal{Y}}(y)]\right\} \\
&\qquad = \frac{1}{2\lambda} \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \pi(x,y)[\ln^2 \pi(y|x) \\
&\qquad\quad - 2\ln \pi(y|x) \ln \pi_{\mathcal{Y}}(y) \\
&\qquad\quad + \ln^2 \pi_{\mathcal{Y}}(y) + (\ln 2) D(\pi(\cdot|x)||\pi_{\mathcal{Y}}) \ln \frac{\pi_{\mathcal{Y}}(y)}{\pi(y|x)}] \\
&\qquad = \frac{1}{2\lambda} \left[\sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \pi(x,y) \ln^2 \frac{\pi_{\mathcal{Y}}(y)}{\pi(y|x)} \right. \\
&\qquad\quad \left. -(\ln 2)^2 \sum_{x\in\mathcal{Y}} t(x) D^2(\pi(\cdot|x)||\pi_{\mathcal{Y}})\right] \\
&\qquad = \sqrt{(2\ln 2)\delta} \left[\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \pi(x,y) \ln^2 \frac{\pi_{\mathcal{Y}}(y)}{\pi(y|x)} \right. \\
&\qquad\quad \left. - \sum_{x\in\mathcal{X}} t(x)(\ln 2)^2 D^2(\pi(\cdot|x)||\pi_{\mathcal{Y}})\right]^{\frac{1}{2}}.
\end{aligned} \tag{B19}$$

Note that $\pi(x,y) = t(x)\pi(y|x)$, $\pi(y|x) = P_{Y|X}(y|x)$, and $\pi_{\mathcal{Y}}(y) = r(y)$. Also, the above argument is valid uniformly for all $t \in \mathcal{P}(\mathcal{X})$ satisfying $t(x) \geq \beta$ for all $x \in \mathcal{X}$.

This completes the proof of Lemma 4. □

Appendix C

In this Appendix, we prove Lemma 5. Fix $x^n \in \mathcal{X}^n$ and type $s \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y})$ according to the lemma's statement. Around s^*, we define the following type set:

$$\mathcal{T}_{s^*} \triangleq \left\{ s \in \mathcal{T}_n(\mathcal{X}\times\mathcal{Y}) : ||s - s^*||_1 \leq \frac{\kappa}{\sqrt{n}}, s_{\mathcal{X}} = t \right\}.$$

We see that the set $B(x^n, s^*)$ is equal to

$$B(x^n, s^*) = \{y^n \in \mathcal{Y}^n : \tau(x^n, y^n) \in \mathcal{T}_{s^*}\}.$$

$$\begin{aligned}
\alpha_x &= t(x) \sum_{y\in\mathcal{Y}} \pi(y|x) \ln \frac{\pi_{\mathcal{Y}}(y)}{\pi(y|x)} \\
&= -(\ln 2) t(x) D(\pi(\cdot|x)||\pi_{\mathcal{Y}})
\end{aligned} \tag{B17}$$

and

$$\begin{aligned}
2\lambda &= \frac{1}{\sqrt{(2\ln 2)\delta}} \sqrt{\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} t(x)\pi(y|x) \left[-\ln \pi(y|x) + \ln \pi_{\mathcal{Y}}(y) - \frac{\alpha_x}{t(x)}\right]^2} \\
&= \frac{1}{\sqrt{(2\ln 2)\delta}} \sqrt{\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} t(x)\pi(y|x) \left[\ln \frac{\pi_{\mathcal{Y}}(y)}{\pi(y|x)} + (\ln 2) D(\pi(\cdot|x)||\pi_{\mathcal{Y}})\right]^2} \\
&= \frac{1}{\sqrt{(2\ln 2)\delta}} \left[\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \pi(x,y) \ln^2 \frac{\pi_{\mathcal{Y}}(y)}{\pi(y|x)} - \sum_{x\in\mathcal{X}} t(x)(\ln 2)^2 D^2(\pi(\cdot|x)||\pi_{\mathcal{Y}})\right]^{\frac{1}{2}}.
\end{aligned} \tag{B18}$$

We pause at this moment to discuss some properties of the set $\mathcal{T}_{s^*}$. When n is sufficiently large, we see that $\mathcal{T}_{s^*}$ is a L1 ball centered at s^* and is thus symmetric with respect to s^*, i.e.

$$\sum_{s\in\mathcal{T}_{s^*}} \frac{s}{|\mathcal{T}_{s^*}|} = s^*. \tag{C1}$$

Furthermore, observe that there are $|\mathcal{X}||\mathcal{Y}| - |\mathcal{X}|$ degrees of freedom to move the entries of s^* at step size $1/n$ to obtain a type in $\mathcal{T}_{s^*}$. This implies that

$$|\mathcal{T}_{s^*}| = 2^{\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log n + O(1)}. \tag{C2}$$

We now count the number of distinct sequences in $B(x^n, s^*)$ as follows:

$$\begin{aligned}
|B(x^n,s^*)| &= \sum_{s\in\mathcal{T}_{s^*}} |\{y^n\in\mathcal{Y}^n : \tau(x^n,y^n)=s\}| \\
&= \sum_{s\in\mathcal{T}_{s^*}} \prod_{x\in\mathcal{X}} \binom{nt(x)}{\{ns(x,y): y\in\mathcal{Y}\}} \\
&= \sum_{s\in\mathcal{T}_{s^*}} \frac{2^{nH(s)-\frac{|\mathcal{X}||\mathcal{Y}|-1}{2}\log n+O(1)}}{2^{nH(t)-\frac{|\mathcal{X}|-1}{2}\log n+O(1)}} \\
&= \sum_{s\in\mathcal{T}_{s^*}} 2^{n[H(s)-H(t)]-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log n+O(1)} \\
&\overset{1)}{\geq} |\mathcal{T}_{s^*}| 2^{n[\sum_{s\in\mathcal{T}_{s^*}}\frac{H(s)}{|\mathcal{T}_{s^*}|}-H(t)]-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log n+O(1)} \\
&\overset{2)}{=} 2^{n[\sum_{s\in\mathcal{T}_{s^*}}\frac{H(s)}{|\mathcal{T}_{s^*}|}-H(t)]+O(1)} \\
&\overset{3)}{=} 2^{n[H(s^*)-H(t)]+O(1)}
\end{aligned} \tag{C3}$$

for sufficiently large n. In the above, the inequality 1) is due to the convexity of the exponential function; the equality 2) is obtained from (C2); and the last equality 3) is argued in Append E.

We next lower bound the probability

$$\Pr\{Y^n\in B(x^n,s^*)|X^n=x^n\}.$$

Let y^n denote a sequence in $B(x^n,s^*)$ constructed above. Then

$$\begin{aligned}
&\Pr\{Y^n=y^n|X^n=x^n\} \\
&\quad= \frac{\Pr\{Y^n=y^n,X^n=x^n\}}{\Pr\{X^n=x^n\}} \\
&\quad= 2^{-n[H(\tau(x^n,y^n))+D(\tau(x^n,y^n)\|P_{XY})]+n[H(t)+D(t\|P_X)]} \\
&\quad= 2^{-n[H(\tau(x^n,y^n))-H(t)+D(\tau(x^n,y^n)\|t\circ P_{Y|X})]}
\end{aligned} \tag{C4}$$

which implies

$$\begin{aligned}
&\Pr\{Y^n\in B(x^n,s^*)|X^n=x^n\} \\
&\quad= \sum_{s\in\mathcal{T}_{s^*}} \Pr\{\tau(x^n,Y^n)=s|X^n=x^n\} \\
&\quad= \sum_{s\in\mathcal{T}_{s^*}} \sum_{y^n\in\mathcal{Y}^n:\tau(x^n,y^n)=s} \Pr\{Y^n=y^n|X^n=x^n\} \\
&\quad= \sum_{s\in\mathcal{T}_{s^*}} |\{y^n\in\mathcal{Y}^n:\tau(x^n,y^n)=s\}| \\
&\qquad\times 2^{-n[H(s)-H(t)+D(s\|t\circ P_{Y|X})]} \\
&\quad\overset{1)}{=} \sum_{s\in\mathcal{T}_{s^*}} 2^{-nD(s\|t\circ P_{Y|X})-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log n+O(1)} \\
&\quad\overset{2)}{\geq} |\mathcal{T}_{s^*}| 2^{-n\sum_{s\in\mathcal{T}_{s^*}}\frac{D(s\|t\circ P_{Y|X})}{|\mathcal{T}_{s^*}|}-\frac{|\mathcal{X}||\mathcal{Y}|-|\mathcal{X}|}{2}\log n+O(1)} \\
&\quad\overset{3)}{=} 2^{-n\sum_{s\in\mathcal{T}_{s^*}}\frac{D(s\|t\circ P_{Y|X})}{|\mathcal{T}_{s^*}|}+O(1)} \\
&\quad= 2^{-nD(s^*\|t\circ P_{Y|X})+O(1)}
\end{aligned} \tag{C5}$$

for sufficiently large n. In the above, the equality 1) follows from an argument similar to that applied in the derivation of (C3) above; the inequality 2) is due to the convexity of the exponential function; the equality 3) follows from (C2); and the last equality follows from an argument similar to that in Appendix E.

It remains to lower bound the cardinality of any subset of $B(x^n,s^*)$ with probability no smaller than $\Pr\{Y^n\in B(x^n,s^*)|X^n=x^n\}-\xi$, where ξ is a real number satisfying $\xi<\Pr\{Y^n\in B(x^n,s^*)|X^n=x^n\}$. Recall that

$$H(p)+D(p\|q)$$

is a linear function of p, where p and q are two distributions defined over the same alphabet. From (C4) and the construction of $B(x^n,s^*)$, it follows that for any $y^n\in B(x^n,s^*)$

$$\begin{aligned}
&\Pr\{Y^n=y^n|X^n=x^n\} \\
&\qquad\leq 2^{-n[H(s^*)-H(t)+D(s^*\|t\circ P_{Y|X})]+c_1\sqrt{n}}
\end{aligned} \tag{C6}$$

where $c_1=-\kappa\log\beta$ is a positive constant. It then follows immediately from (C6) that any subset $B^+(x^n,s^*)$ of $B(x^n,s^*)$ with probability $\Pr\{Y^n\in B(x^n,s^*)|X^n=x^n\}-\xi$ satisfies

$$\begin{aligned}
|B^+(x^n,s^*)| &\geq (\Pr\{Y^n\in B(x^n,s^*)|X^n=x^n\}-\xi) \\
&\quad\times 2^{n[H(s^*)-H(t)+D(s^*\|t\circ P_{Y|X})]-c_1\sqrt{n}}
\end{aligned} \tag{C7}$$

for sufficiently large n. Select $\kappa_0>c_1$ to take care of the $O(1)$ terms in (C3) and (C5). We then get all the desired inequalities of Lemma 5 from (C3), (C5), and (C7). This completes the proof of Lemma 5. □

APPENDIX D

In this Appendix, we prove Lemma 6. Let us partition the set $\mathcal{T}_n(\mathcal{X})$ into a sequence of mutually exclusive subsets $\{\mathcal{S}_k; k\geq 0\}$ defined by

$$\mathcal{S}_k \triangleq \left\{t\in\mathcal{T}_n(\mathcal{X}) : k\frac{1}{\sqrt{n}} \leq \|t-P_X\|_1 < (k+1)\frac{1}{\sqrt{n}}\right\}.$$

For each k, the probability that $\tau(X^n)\in\mathcal{S}_k$ is upper bounded by

$$\begin{aligned}
&\Pr\{\tau(X^n)\in\mathcal{S}_k\} \\
&\quad= \sum_{t\in\mathcal{S}_k} \Pr\{\tau(X^n)=t\} \\
&\quad= \sum_{t\in\mathcal{S}_k} 2^{-nD(t\|P_X)-\frac{|\mathcal{X}|-1}{2}\log n+O(1)} \\
&\quad\overset{1)}{\leq} \sum_{t\in\mathcal{S}_k} 2^{-\frac{n\|t-P_X\|_1^2}{2\ln 2}-\frac{|\mathcal{X}|-1}{2}\log n+O(1)}
\end{aligned}$$

$$\leq \sum_{t\in\mathcal{S}_k} 2^{-\frac{k^2}{2\ln 2}-\frac{|\mathcal{X}|-1}{2}\log n+O(1)}$$
$$\overset{2)}{=} e^{-\frac{k^2}{2}+O(1)}. \tag{D1}$$

In the above, inequality 1) is due to the L1 bound on relative entropy [3, Lemma 12.6.1], i.e.

$$D(p\|q) \geq \frac{1}{2\ln 2}\|p-q\|_1^2$$

and equality 2) follows from observing that

$$|\mathcal{S}_k| \leq 2^{\frac{|\mathcal{X}|-1}{2}\log n}.$$

Using $\{\mathcal{S}_k\}$, we rewrite

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_n(\mathcal{X})} \Pr\{\tau(X^n)=t\}\|t-P_X\|_1 \\
&\quad\leq \frac{1}{\sqrt{n}}\sum_{k=0}^{\infty}\Pr\{\tau(X^n)\in\mathcal{S}_k\}(k+1) \\
&\quad\overset{1)}{\leq} \frac{1}{\sqrt{n}}\sum_{k=0}^{\infty} e^{-\frac{k^2}{2}+O(1)}(k+1) \\
&\quad\leq \frac{\kappa'}{\sqrt{n}}\int_0^{\infty} e^{-\frac{x^2}{2}}(x+1)dx \\
&\quad= \frac{\kappa'(1+\sqrt{\pi/2})}{\sqrt{n}}
\end{aligned} \tag{D2}$$

where κ' is a positive constant. In the above, inequality 1) follows from (D1). Let $\kappa_6 = \kappa'(1+\sqrt{\pi/2})$. This completes the proof of Lemma 6. □

APPENDIX E

In this appendix, we detail the derivation of the equality 3) in (C3). Specifically, we show there exists a constant κ_2, dependent only on β, such that

$$\sum_{s\in\mathcal{T}_{s^*}} \frac{H(s)}{|\mathcal{T}_{s^*}|} \geq H(s^*) - \frac{\kappa_2}{n} \tag{E1}$$

for sufficiently large n.

Regard s and s^* as row vectors for brevity. Expanding $H(s)$ at s^* by using Taylor's series, we get

$$\begin{aligned}
H(s) = H(s^*) &+ \frac{\partial H(s^*)}{\partial s^*}(s-s^*)' \\
&+\frac{1}{2}(s-s^*)\frac{\partial^2 H(s^*)}{\partial (s^*)^2}(s-s^*)' + o\left(\|s-s^*\|_1^2\right)
\end{aligned} \tag{E2}$$

where $(s-s^*)'$ denotes the transpose of $s-s^*$. Since the term $\frac{\partial H(s^*)}{\partial s^*}(s-s^*)'$ is linear with respect to s, it follows from (C1) that

$$\sum_{s\in\mathcal{T}_{s^*}} \frac{1}{|\mathcal{T}_{s^*}|}\frac{\partial H(s^*)}{\partial s^*}(s-s^*)' = 0. \tag{E3}$$

Let m denote the smallest entry in s^*. It is easy to verify that the absolute value of every entry in $\frac{\partial^2 H(s^*)}{\partial (s^*)^2}$ is upper bounded by $\frac{2}{m}$. Consequently,

$$\begin{aligned}
&\frac{1}{2}(s-s^*)\frac{\partial^2 H(s^*)}{\partial (s^*)^2}(s-s^*)' \\
&\quad\geq -\frac{1}{m}\|s-s^*\|_1^2 \\
&\quad\geq -\frac{1}{\beta}\|s-s^*\|_1^2.
\end{aligned} \tag{E4}$$

Combining (E2), (E3), and (E4), we arrive at

$$\begin{aligned}
H(s) &\geq H(s^*) - \frac{1}{\beta}\|s-s^*\|_1^2 + o\left(\|s-s^*\|_1^2\right) \\
&\geq H(s^*) - \frac{\kappa^2}{\beta n} + o\left(\frac{1}{n}\right)
\end{aligned} \tag{E5}$$

where the last inequality follows from the definition of $\mathcal{T}_{s^*}$ in Appendix C. This completes the proof of (E1). □

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the Associate Editor Dr. Ordentlich for their constructive comments that helped improve the presentation of the paper.

REFERENCES

[1] N. Alon and A. Orlitsky, "Source coding and graph entropyies," *IEEE Trans. Inf. Theory*, vol. 42, pp. 1329–1339, 1996.
[2] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "Redundancy rates of Slepian-Wolf coding," in *Proc. Allerton Conf.*, Monticello, IL, USA, 2004.
[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[4] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2505–2523, 1998.
[5] I. Csiszár and J. Körner, "Towards a general theory of source networks," *IEEE Trans. Inf. Theory*, vol. 26, pp. 155–165, 1980.
[6] R. G. Gallager, Source Coding With Side Information and Universal Coding MIT LIDS Tech. Rep. (LIDS-P-937), 1976.
[7] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes," *IEEE Commun. Lett.*, pp. 417–419, 2001.
[8] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed source coding," *Proc. IEEE*, vol. 93, pp. 71–83, 2005.
[9] D.-k. He, L. A. Lastras-Montaño, and E.-h. Yang, "Redundancy of variable rate Slepian-Wolf codes from the decoder's perspective," in *Proc. IEEE Int. Symp. Information Theory (ISIT'07)*, Nice, France, Jun. 2007, pp. 1321–1325.
[10] P. Koulgi, E. Tuncel, S. L. Regunathan, and K. Rose, "On zero-error source coding with decoder side information," *IEEE Trans. Inf. Theory*, vol. 49, pp. 99–111, 2003.
[11] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Trans. Inf. Theory*, vol. 49, pp. 626–643, 2003.
[12] D. Schongberg, K. Ramchandran, and S. S. Pradhan, "Distributed code constructions for the entire Slepian-Wolf rate region for arbitrarily correlated sources," in *Proc. DCC'04*, Snowbird, UT, 2004.
[13] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, pp. 471–480, 1973.
[14] V. Stankovic, A. Liveris, Z. Xiong, and C. Georghiades, "On code design for the general Slepian-Wolf problem and for lossless multiterminal communication networks," *IEEE Trans. Inf. Theory*, submitted for publication.
[15] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. New York: Springer-Verlag, 1978.
[16] E.-H. Yang, D.-K. He, T. Uyematsu, and R. W. Yeung, "Universal multiterminal source coding algorithms with asymptotically zero feedback: Fixed database case," *IEEE Trans. Inf. Theory*, vol. 54, pp. 5575–5590, 2008.
[17] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion—Part one: Known statistics," *IEEE Trans. Inf. Theory*, vol. 43, pp. 71–91, 1997.

[18] H. S. Witsenhausen, "The zero-error side information problem and chromatic numbers," *IEEE Trans. Inf. Theory*, vol. 22, pp. 592–593, 1976.

Da-ke He (S'01–M'06) received the B.S. and M.S. degrees, both in electrical engineering, from Huazhong University of Science and Technology, Wuhan, Hubei, China, in 1993 and 1996, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 2003.

From 1996 to 1998, he worked at Apple Technology China(Zhuhai) as a software engineer. From 2003 to 2004, he worked in the Department of Electrical and Computer Engineering at the University of Waterloo as a postdoctoral research fellow in the Leitch-University of Waterloo Multimedia Communications Laboratory. From 2005 to 2008, he was a Research Staff Member in the Department of Multimedia Technologies at IBM T. J. Watson Research Center in Yorktown Heights, NY. Since 2008, he has been a Technical Manager in Slipstream Data, a subsidiary of Research In Motion, in Waterloo, ON, Canada. His research interests are in source coding theory and algorithm design, multimedia data compression and transmission, multiterminal source coding theory and algorithms, and digital communications.

Luis A. Lastras-Montaño (M'96–SM'06) received the Licenciatura in electronics and digital systems from the Universidad Autónoma de San Luis Potosí, México and the the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY, in 1998 and 2000, respectively.

He has been a Research Staff Member with the IBM T. J. Watson Research Center, Yorktown Heights, NY, since 2000, where he is currently a member of the Memory Systems Department. His academic research interests lie in the areas of information, coding theory and statistical signal processing; his work for IBM is focused on the reliability, performance and architecture of computer memory systems with a special interest on nonvolatile memory technologies.

En-Hui Yang (M'97–SM'00–F'08) was born in Jiangxi, China, on December 26, 1966. He received the B.S. degree in applied mathematics from HuaQiao University, Qianzhou, China, and the Ph.D. degree in mathematics from Nankai University, Tianjin, China, in 1986 and 1991, respectively.

Since June 1997, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, where he is currently a Professor and Canada Research Chair in information theory and multimedia compression. He held a Visiting Professor position at the Chinese University of Hong Kong, Hong Kong, from September 2003 to June 2004; positions of Research Associate and Visiting Scientist at the University of Minnesota, Minneapolis-St. Paul, the University of Bielefeld, Bielefeld, Germany, and the University of Southern California, Los Angeles, from January 1993 to May 1997; and a faculty position (first as an Assistant Professor and then an Associate Professor) at Nankai University, Tianjin, China, from 1991 to 1992. He is the founding Director of the Leitch-University of Waterloo multimedia communications lab, and a Co-Founder of SlipStream Data Inc. (now a subsidiary of Research In Motion). His current research interests are: multimedia compression, multimedia watermarking, multimedia transmission, digital communications, information theory, source and channel coding including distributed source coding and space-time coding, Kolmogorov complexity theory, and applied probability theory and statistics.

Dr. Yang is a recipient of several research awards including the 1992 Tianjin Science and Technology Promotion Award for Young Investigators; the 1992 third Science and Technology Promotion Award of Chinese National Education Committee; the 2000 Ontario Premier's Research Excellence Award, Canada; the 2000 Marsland Award for Research Excellence, University of Waterloo; the 2002 Ontario Distinguished Researcher Award; the prestigious Inaugural (2007) Premier's Catalyst Award for the Innovator of the Year; and the 2007 Ernest C. Manning Award of Distinction, one of the Canada's most prestigious innovation prizes. Products based on his inventions and commercialized by SlipStream received the 2006 Ontario Global Traders Provincial Award and were deployed by over 2200 Service Providers in more than 50 countries, servicing millions of home and wireless subscribers worldwide every day. He served, among many other roles, as a General Co-Chair of the 2008 IEEE International Symposium on Information Theory, a Technical Program Vice-Chair of the 2006 IEEE International Conference on Multimedia & Expo (ICME), the Chair of the award committee for the 2004 Canadian Award in Telecommunications, a Co-Editor of the 2004 Special Issue of the IEEE TRANSACTIONS ON INFORMATION THEORY, a Co-Chair of the 2003 US National Science Foundation (NSF) workshop on the interface of Information Theory and Computer Science, and a Co-Chair of the 2003 Canadian Workshop on Information Theory. He currently also serves as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY. He is a Fellow of the Canadian Academy of Engineering and a Fellow of the Royal Society of Canada: the Academies of Arts, Humanities and Sciences of Canada.

Ashish Jagmohan received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Delhi, in 1999 and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois, at Urbana -Champaign, in 2002 and 2004, respectively.

Since 2004, he has worked at the Departments of Multimedia Technologies and Memory Systems in the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include memory system technologies, video compression, multimedia communication, signal processing, and information theory.

Jun Chen (S'03–M'06) received the B.E. degree with honors in communication 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 2006, 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 T. 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, ON, Canada. He holds the Barber-Gennum Chair in Information Technology. His research interests include information theory, wireless communications, and signal processing.