Near Lossless Source Coding with Side Information at The Decoder: Beyond Conditional Entropy

En-hui Yang
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON N2L 3G1, Canada
Email: ehyang@uwaterloo.ca

Da-ke He
Department of Multimedia Technologies
IBM TJ Watson Research Center
Yorktown Heights, NY 10598, USA
Email: dakehe@us.ibm.com

***Abstract*— In near lossless source coding with decoder only side information, i.e., Slepian-Wolf coding (with one encoder), a source X with finite alphabet $\mathcal{X}$ is first encoded, and then later decoded subject to a small error probability with the help of side information Y with finite alphabet $\mathcal{Y}$ available only to the decoder. The classical result by Slepian and Wolf shows that the minimum average compression rate achievable asymptotically subject to a small error probability constraint for a memoryless pair (X, Y) is given by the conditional entropy $H(X|Y)$. In this paper, we look beyond conditional entropy and investigate the tradeoff between compression rate and decoding error spectrum in Slepian-Wolf coding when the decoding error probability goes to zero exponentially fast. It is shown that when the decoding error probability goes to zero at the speed of $2^{-\delta n}$, where δ is a positive constant and n denotes the source sequences' length, the minimum average compression rate achievable asymptotically is strictly greater than $H(X|Y)$ regardless of how small δ is. More specifically, the minimum average compression rate achievable asymptotically is lower bounded by a quantity called the intrinsic conditional entropy $H_{in}(X|Y, \delta)$, which is strictly greater than $H(X|Y)$, and is also asymptotically achievable for small δ.**

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 convenience, the memoryless sources $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are also referred to simply as the sources X and Y, respectively. In the near lossless source coding of X with decoder only side information Y, the source X is first encoded, and then later decoded subject to a small error probability with the help of the side information Y available only to the decoder. As such, a coding scheme of this type is described by a pair $C_n = (f_n, g_n)$, where f_n, acting as an encoder, encodes $X^n = X_1 X_2 \cdots X_n$ into a binary codeword $f_n(X^n)$, and g_n, acting as a decoder, decodes $f_n(X^n)$ into $g_n(f_n(X^n), Y^n)$, an estimate of X^n, with the help of the side information sequence $Y^n = Y_1 Y_2 \cdots Y_n$. The performance of such a coding scheme is then measured by its compression rate and its decoding error probability $\epsilon_n = \Pr\{X^n \neq g_n(f_n(X^n), Y^n)\}$.

The above source coding paradigm was first proposed and studied by Slepian and Wolf [1]. It was shown [1] that the minimum average compression rate achievable asymptotically subject to a small error probability constraint $\epsilon_n = o(1)$ for a memoryless pair (X, Y) is still given by the conditional entropy $H(X|Y)$ of X given Y, the same rate as in the case where the side information Y is available to both the encoder and decoder. Recently, it was further shown [3] that this result remains valid no matter how fast ϵ_n goes to 0 as long as $-\log \epsilon_n = o(n)$. On the other hand, if the decoding error probability ϵ_n is required to be exactly 0, then Witsenhausen demonstrated [2] that for a memoryless pair (X, Y) whose joint distribution has no zero entries, the minimum average compression rate achievable asymptotically is no longer the conditional entropy $H(X|Y)$, but rather the entropy $H(X)$ of X.

If we interpret

$$\lim_{n \to \infty} \frac{-\log \epsilon_n}{n}$$

as the spectrum of the decoding error probability (i.e., the error exponent), then it is clear that the above results are corresponding to the two far ends of the spectrum. The conditional entropy $H(X|Y)$ is the minimum average compression rate asymptotically achievable at the spectrum of 0, which is one of the first and perhaps one of the best-telling and inspiring results in network information theory. The entropy $H(X)$ is the minimum average compression rate asymptotically achievable at the spectrum of ∞, which is, unfortunately, a negative result in the sense that the side information available only to the decoder does not help at all at this spectrum. What missing is the achievability result in the entire open spectrum other than the end points.

The purpose of this paper is to address the above problem, i.e., the compression rate and error spectrum tradeoff. Specifically, we shall show that at the error spectrum $0 \leq \delta < \infty$, the minimum average compression rate achievable asymptotically is lower bounded by a quantity called the intrinsic conditional entropy $H_{in}(X|Y, \delta)$, which is strictly greater than $H(X|Y)$ regardless of how small $\delta > 0$ is. Furthermore, we shall show that the intrinsic conditional entropy $H_{in}(X|Y, \delta)$ is also asymptotically achievable for small δ. At $\delta = 0$, $H_{in}(X|Y, \delta)$ is equal to $H(X|Y)$. As δ increases and passes a certain point, $H_{in}(X|Y, \delta)$ is flat and equal to $H(X)$.

This work was 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.

II. PRELIMINARIES

In this section, we review the concepts of intrinsic entropy and intrinsic conditional entropy introduced in [3], which will play a fundamental role in our analysis of the compression rate and error spectrum tradeoff.

We first describe the notation to be used throughout this paper. Let $\mathcal{A} = \{a_1, \cdots, a_m\}$ 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 $x_m x_{m+1} \cdots x_n$ as x_m^n, where $m \leq n$ 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

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

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 stands for the base of $\ln$.

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

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

is said to be an n-type if for any $a \in \mathcal{A}$, $t(a) \in \{0, 1/n, 2/n, \cdots, 1\}$. 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 $\tau(x^n) \stackrel{\Delta}{=} (\tau(x^n, a_1), \cdots, \tau(x^n, a_m))$ which is an n-type on $\mathcal{A}$, where $\tau(x^n, a_i) \stackrel{\Delta}{=} \frac{|\{j:\ x_j = a_i\}|}{n}$. 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., $T_{\mathcal{A}}^n(t) \stackrel{\Delta}{=} \{x^n \in \mathcal{A}^n : \tau(x^n) = t\}$.

We now introduce the notion of *intrinsic entropy*. For $\pi \in \mathcal{P}(\mathcal{A} \times \mathcal{B})$ and $\delta \geq 0$, we define the intrinsic δ-entropy of π as

$$H_{in}(\pi, \delta) \stackrel{\Delta}{=} \sup_{\hat{\pi} \in \mathcal{P}(\mathcal{A} \times \mathcal{B}):\ D(\hat{\pi}||\pi) \leq \delta} H(\hat{\pi}). \tag{2.1}$$

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

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

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

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

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

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

$$H_{in|\mathcal{B}}(\pi, \delta) \stackrel{\Delta}{=} \max_{\hat{\pi} \in \mathcal{P}(\mathcal{A} \times \mathcal{B}): D(\hat{\pi}||\pi) \leq \delta} [H(\hat{\pi}) - H(\hat{\pi}_{\mathcal{B}})] \tag{2.2}$$

where $\hat{\pi}_{\mathcal{B}}$ denotes the marginal of $\hat{\pi}$ over $\mathcal{B}$. Similarly, the intrinsic δ-conditional entropy of π with a constrained marginal $\pi_{\mathcal{A}}$ is defined by

$$\begin{aligned} & H_{in|\mathcal{B}}(\pi, \delta|\pi_{\mathcal{A}}) \\ \stackrel{\Delta}{=} & \max_{\hat{\pi} \in \mathcal{P}(\mathcal{A} \times \mathcal{B}): D(\hat{\pi}||\pi) \leq \delta, \pi_{\mathcal{A}} = \hat{\pi}_{\mathcal{A}}} [H(\hat{\pi}) - H(\hat{\pi}_{\mathcal{B}})] \end{aligned} \tag{2.3}$$

From (2.2) and (2.3), it is easy to see that $H(\pi) - H(\pi_{\mathcal{B}}) \leq H_{in|\mathcal{B}}(\pi, \delta|\pi_{\mathcal{A}}) \leq H(\pi_{\mathcal{A}})$, and $H(\pi) - H(\pi_{\mathcal{B}}) \leq H_{in|\mathcal{B}}(\pi, \delta) \leq \log|\mathcal{A}|$. For any $\pi \in \mathcal{P}^+(\mathcal{A} \times \mathcal{B})$, let $\delta_{max}(\pi|\pi_{\mathcal{A}}) \stackrel{\Delta}{=} \min\{\delta : H_{in|\mathcal{B}}(\pi, \delta|\pi_{\mathcal{A}}) = H(\pi_{\mathcal{A}})\}$, and $\delta'_{max}(\pi) \stackrel{\Delta}{=} \min\{\delta : H_{in|\mathcal{B}}(\pi, \delta) = \log|\mathcal{A}|\}$. The following lemma is proved in [3].

Lemma 1: The intrinsic conditional entropy $H_{in|\mathcal{B}}(\pi, \delta|\pi_{\mathcal{A}})$ ($H_{in|\mathcal{B}}(\pi, \delta)$, 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).

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 marginal 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. Applying the concept of intrinsic conditional entropy to (X, Y), we have the following result, the proof of which can also be found in [3].

Lemma 2 (intrinsic conditional vs classical conditional): Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$ and $I(X; Y) > 0$. 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 $\delta \leq \Delta_2$ and any distribution $t \in \mathcal{P}(\mathcal{X})$ satisfying $t(x) \geq \beta$ for all $x \in \mathcal{X}$,

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

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

$$\begin{aligned} d_v(t) \quad = \quad & \Big\{ 2 \ln 2 \Big[\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} t(x) P_{Y|X}(y|x) \log^2 \frac{r(y)}{P_{Y|X}(y|x)} - \\ & \sum_{x \in \mathcal{X}} t(x) D^2(P_{Y|X}(\cdot|x)||r) \Big] \Big\}^{1/2} . \end{aligned}$$

In the following, we will make use of intrinsic conditional entropy, and the above lemmas to investigate the compression rate and error spectrum tradeoff. For convenience, we shall sometimes write $H_{in|\mathcal{Y}}(P_{XY}, \delta|P_X)$ simply as $H_{in}(X|Y, \delta)$.

III. LOWER BOUND

In this section, we establish a lower bound on the compression rate of Slepian-Wolf coding of X with decoder side information Y under the condition that the decoding error probability goes to zero exponentially fast. We begin with the formal definition of a Slepian-Wolf code.

Let $\mathcal{I}$ denote a set of finite binary codewords satisfying the prefix condition. An order n Slepian-Wolf 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 Slepian-Wolf coding. The order n Slepian-Wolf 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 Slepian-Wolf codes includes that of all (order n) fixed-rate Slepian-Wolf 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

$$r(C_n) \stackrel{\Delta}{=} \frac{1}{n} \mathrm{E}|f_n(X^n)|,$$

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

$$P_e(C_n) \stackrel{\Delta}{=} \Pr\{X^n \neq \hat{X}^n\}.$$

In [1] and [3], it was shown that at the error spectrum of 0, $r_{C_n}(X^n)$ can be made arbitrarily close to $H(X|Y)$. Below we will examine what will happen for positive error spectra, i.e., when

$$\lim_{n\to\infty} \frac{-\log P_e(C_n)}{n} = \delta > 0\ .$$

We have the following result.

Theorem 1: Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$ and $I(X;Y) > 0$. Then for any δ satisfying $0 < \delta < \Delta$, where $\Delta \stackrel{\Delta}{=} \delta_{max}(P_{XY}|P_Y)$ is defined in Section II, there exists an integer N such that for any $n \geq N$, and for any order n variable rate code $C_n = (f_n, g_n)$ with error probability $P_e(C_n)$,

$$r(C_n) \geq H_{in}(X|Y,\delta) - O\left(\frac{1}{\sqrt{n}}\right)$$

whenever $P_e(C_n) \leq 2^{-\delta n}$.

Our proof of Theorem 1 is based on a type and sphere packing argument similar to that used in the proof of Theorem 1 in [3], and the convexity of $H_{in}(X|Y, \frac{-\log\epsilon}{n})$ as a function of $0 < \epsilon < 1$. Thus, we shall leave the complete proof in the full paper [5], and instead focus on proving that $H_{in}(X|Y, \frac{-\log\epsilon}{n})$ is convex in ϵ for sufficiently large n in Appendix A.

Remark 1: It is instructive to compare Theorem 1 with the previous achievable rate result for Slepian-Wolf coding. As mentioned early, it was shown in [1], [3] that one just needs $H(X|Y)$ bits per symbol to code X with decoder side information Y while maintaining a diminishing decoding error probability. According to Theorem 1, however, this widely accepted statement is not valid anymore when the decoding error probability goes to 0 exponentially. Specifically, in view of Lemma 1, no matter how small $\delta > 0$ is, $H_{in}(X|Y,\delta)$ is always strictly greater than $H(X|Y)$. Further from Lemma 2, the difference between them is in the order of $\sqrt{\delta}$ for any $0 < \delta < \Delta_2$.

IV. UPPER BOUND

In this section, we show that the lower bound established in Section III is tight for small δ by providing a matching upper bound.

Theorem 2: Assume $P_{XY} \in \mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$ and $I(X;Y) > 0$. Let δ be a small positive constant such that

$$\frac{\partial H_{in}(X|Y,\delta)}{\partial \delta} > 1. \tag{4.1}$$

Then there exists a sequence of Slepian-Wolf codes $\{C_n\}_{n=1}^{\infty}$ with $P_e(C_n) \leq 2^{-\delta n}$ such that for sufficiently large n,

$$r_{C_n} \leq H(X|Y,\delta) + O\left(\frac{\log n}{n}\right). \tag{4.2}$$

Proof of Theorem 2: Let $\epsilon_n = 2^{-\delta n}$. Fix a type $s \in \mathcal{T}^n(\mathcal{X} \times \mathcal{Y})$. For every sequence $x^n \in T^n_{\mathcal{X}}(s_{\mathcal{X}})$, define

$$A_{x^n}(s) \stackrel{\Delta}{=} \{y^n \in \mathcal{Y}^n : \tau(x^n, y^n) = s\},$$

and for every $y^n \in T^n_{\mathcal{Y}}(s_{\mathcal{Y}})$, define

$$B_{y^n}(s) \stackrel{\Delta}{=} \{x^n \in \mathcal{X}^n : \tau(x^n, y^n) = s\}.$$

In the following, we construct a Slepian-Wolf code C_n. For brevity, let us fix a type $t \in \mathcal{T}_n(\mathcal{X})$ such that $||t - P_X||_1 \leq \kappa_0 \sqrt{\frac{\log n}{n}}$, where κ_0 is a constant such that

$$\Pr\{||\tau(X^n) - P_X||_1 \geq \kappa_0 \sqrt{\frac{\log n}{n}}\} \leq \frac{1}{n^2}.$$

Let $nR(t)$ denote a positive integer depending only on t and n.

Encoding: C_n randomly puts each sequence $x^n \in T^n_{\mathcal{X}}(t)$ into one of $2^{nR(t)}$ bins with probability $2^{-nR(t)}$. In the following, we shall use $f_t : T^n_{\mathcal{X}}(t) \to \{0, 1, \cdots, 2^{nR(t)} - 1\}$ to associate each sequence $x^n \in T^n_{\mathcal{X}}(t)$ with a bin index. To encode x^n, C_n encodes the type t by using $O(\log n)$ bits, and the bin index $f_t(x^n)$ by using $nR(t)$ bits.

Decoding: Given the bin index $f_t(x^n)$ and a side information sequence Y^n, the decoder of C_n uses the maximum-likelihood decoding rule:

$$\hat{x}^n \stackrel{\Delta}{=} \arg\min_{\substack{\tilde{x}^n \in T^n_{\mathcal{X}}(t):\\ f_t(\tilde{x}^n) = f_t(x^n)}} \Big[H(\tau(\tilde{x}^n, Y^n) + D(\tau(\tilde{x}^n, Y^n)||t \circ P_{Y|X}) \Big].$$

For sequences x^n with $||\tau(x^n) - P_X|| \geq \kappa_0 \sqrt{\frac{\log n}{n}}$, the code C_n sends the type $\tau(x^n)$ along with the binary description of

x^n directly to the decoder; in such a case, there will be no decoding error.

We now analyze the decoding error probability of the above code C_n. Define

$$\beta_n = \frac{-\log \epsilon_n + \kappa_1 \log n}{n}, \tag{4.3}$$

and

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

where $\kappa_1 > 4$ is a positive constant selected so that

$$\Pr\{\tau(x^n, Y^n) \notin \mathcal{S}^+(t, \beta_n) | X^n = x^n\} \le \frac{\epsilon_n}{2}. \tag{4.4}$$

It is not hard to see that κ_1 depends only upon $|\mathcal{X}|$ and $|\mathcal{Y}|$. For each sequence $x^n \in T^n_{\mathcal{X}}(t)$ and for each type $s \in \mathcal{S}^+(t, \beta_n)$, we further define

$$\begin{aligned} \mathcal{S}^+(t, \beta_n, s) \triangleq &\Big\{ s' \in \mathcal{S}^+(t, \beta_n) : s'_{\mathcal{Y}} = s_{\mathcal{Y}} \text{ and} \\ H(s') + D(s' || t \circ P_{Y|X}) &\le H(s) + D(s || t \circ P_{Y|X}) \Big\}. \end{aligned}$$

Suppose that $X^n = x^n$, $Y^n = y^n$, and $s = \tau(x^n, y^n)$ satisfying $s \in \mathcal{S}^+(t, \beta_n)$. Then it follows from the construction of C_n that

$$\begin{aligned} &\Pr\{\text{Error} | X^n = x^n, Y^n = y^n\} \\ = \quad &\sum_{s' \in \mathcal{S}^+(t, \beta_n, s)} \Pr\Big\{ \text{there exists a sequence } \tilde{x}^n \ne x^n \\ &\qquad \text{such that } y^n \in A_{\tilde{x}^n}(s') \text{ and } f_t(\tilde{x}^n) = f_t(x^n) \Big\} \\ \le \quad &\sum_{s' \in \mathcal{S}^+(t, \beta_n, s)} |B_{y^n}(s')| 2^{-nR(t)} \\ \le \quad &|B_{y^n}(s^*_{\beta_n, s})| 2^{-nR(t) + (|\mathcal{X}||\mathcal{Y}| - |\mathcal{X}| - |\mathcal{Y}| + 1) \log n}, \end{aligned} \tag{4.5}$$

where $s^*_{\beta_n, s} \triangleq \arg\max_{s' \in \mathcal{S}^+(t, \beta_n, s)} H(s')$. Observe that Equation (4.5) holds for all (x^n, y^n) satisfying $\tau(x^n, y^n) = s$. Using (4.5), we now calculate the decoding error as follows.

$$\begin{aligned} &\Pr\{\text{Error} | X^n = x^n\} \\ \overset{1)}{=} \quad &\frac{\epsilon_n}{2} + \sum_{s \in \mathcal{S}^+(t, \beta_n)} \Pr\{\tau(X^n, Y^n) = s | X^n = x^n\} \times \\ &\quad \Pr\{\text{Error} | X^n = x^n, \tau(X^n, Y^n) = s\} \\ \le \quad &\frac{\epsilon_n}{2} + \sum_{s \in \mathcal{S}^+(t, \beta_n)} |B_{y^n}(s^*_{\beta_n, s})| \times \\ &\quad 2^{-n[D(s || t \circ P_{Y|X}) + R(t)] + \frac{|\mathcal{X}||\mathcal{Y}| - |\mathcal{X}| - 2|\mathcal{Y}| + 2}{2} \log n + O(1)} \\ = \quad &\frac{\epsilon_n}{2} + \sum_{s \in \mathcal{S}^+(t, \beta_n)} 2^{n[H(s^*_{\beta_n, s}) - H(s_{\mathcal{Y}}) - R(t) - D(s || t \circ P_{Y|X})]} \times \\ &\quad 2^{-\frac{|\mathcal{X}| + |\mathcal{Y}| - 2}{2} \log n + O(1)} \\ \overset{2)}{\le} \quad &\frac{\epsilon_n}{2} + \epsilon_n 2^{-\frac{|\mathcal{X}| + |\mathcal{Y}| + 2\kappa_1 - 2}{2} \log n + O(1)} \times \\ &\sum_{s \in \mathcal{S}^+(t, \beta_n)} 2^{n[H(s^*_{\beta_n, s}) - H(s_{\mathcal{Y}}) - R(t) - D(s^*_{\beta_n, s} || t \circ P_{Y|X}) + \beta_n]}. \end{aligned} \tag{4.6}$$

In the above, the inequality 1) is due to (4.4) and the inequality 2) follows from the definitions of $s^*_{\beta_n, s}$ and $\mathcal{S}^+(t, \beta_n, s)$ above. Select now that

$$R(t) = H_{in|\mathcal{Y}}(t \circ P_{Y|X}, \beta_n | t) + |\mathcal{X}||\mathcal{Y}| \frac{\log n}{n}.$$

In view of the definition of intrinsic conditional entropy and that $||t - P_X||_1 \le \kappa_0 \sqrt{\log n / n}$, it follows from (4.1) that

$$\frac{\partial H_{in|\mathcal{Y}}(P_{XY}, \delta | t)}{\partial \delta} \Big|_{\delta = \beta_n} > 1, \tag{4.7}$$

for sufficiently large n. Since $H_{in|\mathcal{Y}}(P_{XY}, \delta | t)$ as a function of δ is concave, we then have

$$\begin{aligned} &H(s^*_{\beta_n, s}) - H(s_{\mathcal{Y}}) - R(t) - D(s^*_{\beta_n, s} || t \circ P_{Y|X}) + \beta_n \\ < \quad &-|\mathcal{X}||\mathcal{Y}| \frac{\log n}{n} \end{aligned} \tag{4.8}$$

for sufficiently large n. Putting (4.8) back into (4.6), we get

$$\begin{aligned} &\Pr\{\text{Error} | X^n = x^n\} \\ \le \quad &\frac{\epsilon_n}{2} + \\ &\epsilon_n \sum_{s \in \mathcal{S}^+(t, \beta_n)} 2^{-(|\mathcal{X}||\mathcal{Y}| + \kappa_1) \log n - \frac{|\mathcal{X}| + |\mathcal{Y}| - 2}{2} \log n + O(1)} \\ \le \quad &\frac{\epsilon_n}{2} + \epsilon_n 2^{-\frac{3|\mathcal{X}| + 3|\mathcal{Y}| + 2\kappa_1 - 4}{2} \log n + O(1)} \\ \le \quad &\epsilon_n, \end{aligned} \tag{4.9}$$

whenever n is sufficiently large.

To complete the proof, we calculate the rate of C_n as follows:

$$\begin{aligned} r(C_n) \le &\; O\left(\frac{\log n}{n}\right) + \frac{1}{n^2} \log |\mathcal{X}| + \\ &\sum_{t \in \mathcal{T}_n(\mathcal{X}) : ||t - P_X||_1 \le \kappa_0 \sqrt{\log n / n}} \Pr\{\tau(X^n) = t\} R(t) \\ \le \quad &O\left(\frac{\log n}{n}\right) + \frac{\log |\mathcal{X}|}{n^2} \\ &\sum_{t \in \mathcal{T}_n(\mathcal{X}) : ||t - P_X||_1 < \kappa_0 \sqrt{\log n / n}} \Pr\{\tau(X^n) = t\} \times \\ &\quad H_{in|\mathcal{Y}}(t \circ P_{Y|X}, \beta_n | t) \\ \le \quad &O\left(\frac{\log n}{n}\right) + H_{in|\mathcal{Y}}(P_{XY}, \beta_n | P_X) \\ = \quad &H_{in|\mathcal{Y}}(P_{XY}, \delta | P_X) + O\left(\frac{\log n}{n}\right), \end{aligned} \tag{4.10}$$

where the last inequality is due to Lemma 1.

Combining (4.9) with (4.10) now implies that there exists an order n deterministic Slepian-Wolf code with the decoding error probability less than or equal to ϵ_n and the average compression rate upper bounded by (4.10). This completes the proof of Theorem 2.

Recall that the intrinsic conditional entropy $H_{in}(X|Y, \delta)$ is a concave function of δ according to Lemma 1, in other words, $\frac{H_{in}(X|Y, \delta)}{\partial \delta}$ is non-decreasing as δ increases. This fact, together with Theorem 2, implies that $H_{in}(X|Y, \delta)$ is achievable asymptotically for small δ.

V. CONCLUSIONS

We have revisited Slepian-Wolf coding (with one encoder) from the new perspective of error spectrum for jointly memoryless source-side information pair (X, Y) whose joint distribution has no zero entries. In this perspective, the classical Slepian-Wolf result [1] and the zero-error Witsenhausen result [2] correspond to the two far ends where the error spectra are 0 and ∞, respectively. In this paper we have shown that at any error spectrum $0 \leq \delta < \infty$, the minimum average compression rate achievable asymptotically is lower bounded by the intrinsic conditional entropy $H_{in}(X|Y, \delta)$. At $\delta = 0$, the quantity $H_{in}(X|Y, \delta)$ is equal to $H(X|Y)$. As δ increases and passes a certain point, $H_{in}(X|Y, \delta)$ is flat and equal to $H(X)$. For any δ in between, $H_{in}(X|Y, \delta)$ is strictly greater than $H(X|Y)$. Furthermore, we have show that this lower bound is tight for small δ in the sense that the intrinsic conditional entropy $H_{in}(X|Y, \delta)$ is indeed asymptotically achievable.

APPENDIX A

In this appendix, we investigate the convexity of $H_{in|\mathcal{Y}}(\pi, \frac{-\log \epsilon}{n})$ as a function of ϵ, where π denotes a fixed distribution in $\mathcal{P}^+(\mathcal{X} \times \mathcal{Y})$. For convenience, we assume all logarithms have base e in this appendix; as such, all involved quantities will be scaled by a common factor. It has been shown in [3] that $H_{in|\mathcal{Y}}(\pi, \delta)$ is a concave function of (π, δ) and is strictly increasing in $[0, \delta_{max}(\pi)]$ for any fixed π. In the following, we shall utilize this result to prove that $H_{in|\mathcal{Y}}(\pi, \frac{-\log \epsilon}{n})$ is convex in ϵ whenever n is large enough. To this end, let us calculate the second order derivative of $H_{in|\mathcal{Y}}(\pi, \frac{-\log \epsilon}{n})$ with respect to ϵ as follows. For brevity, we shall use $f(\delta)$ to denote $H_{in|\mathcal{Y}}(\pi, \delta)$.

$$\begin{aligned}
\frac{d^2 f(\frac{-\log \epsilon}{n})}{d\epsilon^2} &= f''(\delta)\left(\frac{d\delta}{d\epsilon}\right)^2 + f'(\delta)\left(\frac{d^2\delta}{d\epsilon^2}\right) \\
&= f''(\delta)\left(\frac{1}{n\epsilon}\right)^2 + f'(\delta)\left(\frac{1}{n\epsilon^2}\right) \\
&= \frac{1}{n\epsilon^2}\left[\frac{1}{n}f''(\delta) + f'(\delta)\right]. \qquad \text{(A.1)}
\end{aligned}$$

For our purpose, we are interested in $\delta = \frac{-\log \epsilon}{n}$ lying in $(0, \delta_{max}(\pi))$. Observe that in (A.1), $f''(\delta) < 0$ because $f(\delta)$ is concave, and $f'(\delta) > 0$ because $f(\delta)$ is strictly increasing. In order to show that $f(\epsilon)$ is convex, it suffices to lower bound the sum inside the brackets from 0. To this end, let us first derive an expression of $f''(\delta)$.

In view of the definition of intrinsic conditional entropy $H_{in|\mathcal{Y}}(\pi, \delta)$, we see that $H_{in|\mathcal{Y}}(\pi, \delta) = H(s^*)$ where s^* is a solution to the following maximization problem:

$$\max H(s) - H(s_{\mathcal{Y}}) - \lambda D(s||\pi) \qquad \text{(A.2)}$$

where s denotes a probability distribution over $\mathcal{X} \times \mathcal{Y}$. In the above, λ is the standard Lagrange multiplier that is equal to $f'(\delta)$. Solving (A.2) by fixing λ leads to

$$s^*(x, y) = \frac{s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}} \pi^{\frac{\lambda}{1+\lambda}}(x, y)}{\sum_{y' \in \mathcal{Y}} \Phi_{y'}} \quad \text{for } (x, y) \in \mathcal{X} \times \mathcal{Y}, \text{(A.3)}$$

where for any $y \in \mathcal{Y}$

$$\Phi_y \overset{\Delta}{=} s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}} \sum_{x \in \mathcal{X}} \pi^{\frac{\lambda}{1+\lambda}}(x, y) \qquad \text{(A.4)}$$

For brevity, let us write $\sum_{y \in \mathcal{Y}} \Phi_y$ simply as Φ. Regard $s^*(x, y)$ as a function of λ. We can calculate

$$\begin{aligned}
&\frac{ds^*(x, y)}{d\lambda} \\
&= \frac{s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}} \pi^{\frac{\lambda}{1+\lambda}}(x, y)}{\Phi}\Big[\frac{-1}{(1+\lambda)^2}\ln s^*_{\mathcal{Y}}(y) + \\
&\quad \frac{1}{(1+\lambda)s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} + \frac{1}{(1+\lambda)^2}\ln \pi(x, y)\Big] - \\
&\quad \frac{s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}} \pi^{\frac{\lambda}{1+\lambda}}(x, y)}{\Phi^2}\Big[\sum_{y' \in \mathcal{Y}} \frac{d\Phi_{y'}}{d\lambda}\Big] \\
&= \frac{s^*(x, y)}{1+\lambda}\Big[\frac{-1}{1+\lambda}\ln \frac{s^*_{\mathcal{Y}}(y)}{\pi(x, y)} + \frac{1}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} - \\
&\quad \frac{1+\lambda}{\Phi}\sum_{y' \in \mathcal{Y}} \frac{d\Phi_{y'}}{d\lambda}\Big] \qquad \text{(A.5)}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{d\Phi_y}{d\lambda} \\
&= s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}}\Big[\frac{-1}{(1+\lambda)^2}\ln s^*_{\mathcal{Y}}(y) + \\
&\quad \frac{1}{(1+\lambda)s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\Big]\sum_{x \in \mathcal{X}} \pi^{\frac{\lambda}{1+\lambda}}(x, y) + \\
&\quad s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}}\sum_{x \in \mathcal{X}} \frac{\pi(x, y)^{\frac{\lambda}{1+\lambda}}}{(1+\lambda)^2}\ln \pi(x, y) \\
&= \frac{\Phi_y}{1+\lambda}\Big[\frac{-1}{1+\lambda}\ln s^*_{\mathcal{Y}}(y) + \\
&\quad \frac{1}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\Big] + s^*_{\mathcal{Y}}(y)^{\frac{1}{1+\lambda}}\sum_{x \in \mathcal{X}} \frac{\pi(x, y)^{\frac{\lambda}{1+\lambda}}}{(1+\lambda)^2}\ln \pi(x, y) \\
&= \frac{\Phi}{1+\lambda}\Big\{\frac{\Phi_y}{\Phi}\Big[\frac{-1}{1+\lambda}\ln s^*_{\mathcal{Y}}(y) + \frac{1}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\Big] + \\
&\quad \sum_{x \in \mathcal{X}} \frac{s^*(x, y)}{1+\lambda}\ln \pi(x, y)\Big\} \\
&= \frac{\Phi}{1+\lambda}\Big[\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} - \sum_{x \in \mathcal{X}} \frac{s^*(x, y)}{1+\lambda}\ln \frac{s^*(y)}{\pi(x, y)}\Big]. \qquad \text{(A.6)}
\end{aligned}$$

Putting (A.6) into (A.5), we get

$$\begin{aligned}
&\frac{ds^*(x, y)}{d\lambda} \\
&= \frac{s^*(x, y)}{1+\lambda}\Big[\frac{-1}{1+\lambda}\ln \frac{s^*_{\mathcal{Y}}(y)}{\pi(x, y)} + \frac{1}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} - \\
&\quad \sum_{y' \in \mathcal{Y}} \frac{ds^*_{\mathcal{Y}}(y')}{d\lambda} + \ln \Phi + \delta\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{s^*(x,y)}{1+\lambda}\Big[-\ln\frac{s^*(x,y)}{\pi(x,y)} + \frac{1}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} - \\
&\qquad \sum_{y'\in\mathcal{Y}}\frac{ds^*_{\mathcal{Y}}(y')}{d\lambda} + \delta\Big] \qquad \text{(A.7)}
\end{aligned}
$$

Note that

$$\delta = D(s^*||\pi) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} s^*(x,y)\ln\frac{s^*(x,y)}{\pi(x,y)}.$$

It follows that

$$
\begin{aligned}
&\frac{d\delta}{d\lambda} = \sum_{x,y}\Big[\ln\frac{s^*(x,y)}{\pi(x,y)}+1\Big]\frac{ds^*(x,y)}{d\lambda}\\
&= \sum_{x,y}\frac{ds^*(x,y)}{d\lambda}\ln\frac{s^*(x,y)}{\pi(x,y)} + \sum_{x,y}\frac{ds^*(x,y)}{d\lambda}\\
&= \sum_{x,y}\frac{s^*(x,y)}{1+\lambda}\Big[-\ln^2\frac{s^*(x,y)}{\pi(x,y)} + \\
&\qquad \frac{1}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\ln\frac{s^*(x,y)}{\pi(x,y)} - \\
&\qquad \ln\frac{s^*(x,y)}{\pi(x,y)}\sum_{y'\in\mathcal{Y}}\frac{ds^*_{\mathcal{Y}}(y')}{d\lambda} + \delta\ln\frac{s^*(x,y)}{\pi(x,y)}\Big]\\
&= \frac{1}{1+\lambda}\Big[-\sum_{x,y}s^*(x,y)\ln^2\frac{s^*(x,y)}{\pi(x,y)} + \delta^2 + \\
&\qquad \sum_{x,y}\frac{s^*(x,y)}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\ln\frac{s^*(x,y)}{\pi(x,y)} - \delta\sum_{y'\in\mathcal{Y}}\frac{ds^*_{\mathcal{Y}}(y')}{d\lambda}\Big]\\
&= \frac{1}{1+\lambda}\Big[-\sum_{x,y}s^*(x,y)\big(\ln\frac{s^*(x,y)}{\pi(x,y)} - \delta\big)^2 + \\
&\qquad \sum_{x,y}\frac{s^*(x,y)}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\ln\frac{s^*(x,y)}{\pi(x,y)} - \\
&\qquad \delta\sum_{y'\in\mathcal{Y}}\frac{ds^*_{\mathcal{Y}}(y')}{d\lambda}\Big]. \qquad \text{(A.8)}
\end{aligned}
$$

Verify from (A.7) and the fact that $s^*_{\mathcal{Y}}(y) = \sum_{x\in\mathcal{X}} s^*(x,y)$ that for any $\lambda < \infty$,

$$\sum_{y\in\mathcal{Y}}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} = 0, \qquad \text{(A.9)}$$

and for any $y \in \mathcal{Y}$

$$\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda} = -\sum_{x\in\mathcal{X}}\frac{s^*(x,y)}{\lambda}\ln\frac{s^*(x,y)}{\pi(x,y)} + \frac{\delta}{\lambda}s^*_{\mathcal{Y}}(y), \qquad \text{(A.10)}$$

which in turn implies

$$
\begin{aligned}
&\sum_{x,y}\frac{s^*(x,y)}{s^*_{\mathcal{Y}}(y)}\frac{ds^*_{\mathcal{Y}}(y)}{d\lambda}\ln\frac{s^*(x,y)}{\pi(x,y)}\\
&= \sum_{x,y}\frac{s^*(x,y)}{s^*_{\mathcal{Y}}(y)}\ln\frac{s^*(x,y)}{\pi(x,y)}\Big[-\sum_{x'\in\mathcal{X}}\frac{s^*(x',y)}{\lambda}\ln\frac{s^*(x',y)}{\pi(x',y)} + \\
&\qquad \frac{\delta}{\lambda}s^*_{\mathcal{Y}}(y)\Big]\\
&= \frac{1}{\lambda}\sum_{x,y}s^*(x,y)\ln\frac{s^*(x,y)}{\pi(x,y)} \times \\
&\qquad \Big[-\sum_{x'\in\mathcal{X}}\frac{s^*(x',y)}{s^*_{\mathcal{Y}}(y)}\ln\frac{s^*(x',y)}{\pi(x',y)} + \delta\Big]\\
&= \frac{1}{\lambda}\sum_{y}s^*_{\mathcal{Y}}(y)(-\delta_y^2 + \delta_y\delta)\\
&\overset{1)}{=} -\frac{1}{\lambda}\sum_{y}s^*_{\mathcal{Y}}(y)(\delta_y - \delta)^2, \qquad \text{(A.11)}
\end{aligned}
$$

where for any $y \in \mathcal{Y}$,

$$\delta_y \triangleq \sum_{x\in\mathcal{X}}\frac{s^*_{x,y}}{s^*_{\mathcal{Y}}(y)}\ln\frac{s^*(x,y)}{\pi(x,y)}.$$

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

$$\sum_{y\in\mathcal{Y}} s^*_{\mathcal{Y}}\delta_y = \delta.$$

Combining (A.8), (A.9), and (A.11), we arrive at

$$
\begin{aligned}
&f''(\delta) = \frac{d\lambda}{d\delta}\\
&= (1+\lambda)/\Big[-\sum_{x,y}s^*(x,y)\big(\ln\frac{s^*(x,y)}{\pi(x,y)} - \\
&\qquad \delta\big)^2 - \frac{1}{\lambda}\sum_{y}s^*_{\mathcal{Y}}(y)(\delta_y-\delta)^2\Big]\\
&\geq \frac{1+\lambda}{-\sum_{x,y}s^*(x,y)\big(\ln\frac{s^*(x,y)}{\pi(x,y)} - \delta\big)^2},
\end{aligned}
$$

and in turn,

$$
\begin{aligned}
&f'(\delta) + \frac{1}{n}f''(\delta)\\
&\geq \lambda - \frac{1+\lambda}{n\sum_{x,y}s^*(x,y)\big(\ln\frac{s^*(x,y)}{\pi(x,y)} - \delta\big)^2}. \qquad \text{(A.12)}
\end{aligned}
$$

Under the assumption that $\pi_{\mathcal{X}}$ is not uniform, one can lower bound $\sum_{x,y}s^*(x,y)\big(\ln\frac{s^*(x,y)}{\pi(x,y)} - \delta\big)^2$ from 0 whenever $\lambda < \infty$. Consequently, equations (A.1) and (A.12) imply that for any $\delta^* \in (0, \delta'_{max}(\pi))$, there exists a positive integer N such that for all $n \geq N$, $H_{in|\mathcal{Y}}(\pi, \frac{-\log\epsilon}{n})$ is convex in ϵ whenever $\frac{-\log\epsilon}{n} \in (\delta^*, \delta'_{max}(\pi) - \delta^*)$.

Similarly, when $H(\pi_{\mathcal{X}}) + H(\pi_{\mathcal{Y}}) - H(\pi) > 0$, one can show that that for any $\delta^* \in (0, \delta_{max}(\pi|\pi_{\mathcal{X}}))$, there exists a positive integer N such that for all $n \geq N$, $H_{in|\mathcal{Y}}(\pi, \frac{-\log\epsilon}{n}|\pi_{\mathcal{X}})$ is convex in ϵ whenever $\frac{-\log\epsilon}{n} \in (\delta^*, \delta_{max}(\pi|\pi_{\mathcal{X}}) - \delta^*)$.

ACKNOWLEDGMENT

The second author would like to acknowledge Luis A. Lastras-Montaño and Ashish Jagmohan at IBM TJ Watson Research Center, and Jun Chen at McMaster University for their constructive criticism and always enlightening discussions.

REFERENCES

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, 1973.

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

[3] D.-K. He, L. A. Lastras, E.-H. Yang, A. Jagmohan, and J. Chen, "On the redundancy of Slepian-Wolf coding," to be submitted to *IEEE Trans. Inform. Theory*, 2008.

[4] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505-2523, 1998.

[5] E.-H. Yang and D.-K. He, "Near Lossless Source Coding with Side Information at The Decoder: Beyond Conditional Entropy," to be submitted to *IEEE Trans. Inform. Theory*, 2008.