Non-Asymptotic Performance of Symmetric Slepian-Wolf Coding

Shriram Sarvotham, Dror Baron, and Richard Baraniuk[1]
Department of Electrical and Computer Engineering
Rice University, 6100 Main Street, Houston, Texas 77005
e-mail: {shri, drorb, richb}@rice.edu

Abstract — **Like most results in information theory, the classical Slepian-Wolf theorem is asymptotic. We characterize how quickly the limits put forth by information theory can be reached. Our contribution is two-fold. First, we investigate the non-asymptotic regime for two binary sources that generate sequences with a symmetric joint distribution. Second, we provide tight converse and achievable bounds and also show how to construct codes for any rate pair within the feasible rate region. The key result is that the feasible rate region for the non-asymptotic regime is obtained by translating the corresponding rate region for the asymptotic regime.**

I. Introduction

Distributed communication and processing are revolutionizing today's communication systems, replacing more traditional centralized architectures. Centralized architectures consume far greater communication resources and power. Distributed processing, on the other hand, is scalable, load-balanced, robust, and well suited to a resource-constrained environment. A sample application is to compress the data generated in a sensor network with minimal (or no) inter-sensor communication [1–3].

The classical Slepian-Wolf theorem [4] provides a theoretical foundation to study distributed compression. Like most results in information theory, the Slepian-Wolf theorem is asymptotic. Yet practical code design needs to consider finite length codewords. In this work, we probe the non-asymptotic regime and characterize how quickly we can approach the limits on the efficiency of communication systems put forth by information theory. Our previous work [5] has studied the non-asymptotic regime for a specific setup in which side information passes through an independent correlation channel. For that setup, we studied non-asymptotic aspects of coding when complete side information is available at the decoder. Our goal is to extend these results to an arbitrary number of sources, arbitrary alphabet size, and more general distributions. To this end, the work presented in this paper is for a symmetric two source problem.

Our contribution is the determination of achievable and converse bounds for Slepian-Wolf coding for two symmetric binary sources. Consider two correlated length-n sequences x and y, where each sequence is independent and identically distributed (iid) Bernoulli with parameter 0.5; for brevity we denote this distribution by Bernoulli(0.5). The correlation structure is also memoryless in the sense that $z = x \oplus y$ is iid with Bernoulli(p). Let R_x and R_y be the coding rates for the sources X and Y. For this setup, we characterize the rate region (R_x, R_y) for non-asymptotic Slepian-Wolf coding where the probability of codeword error is bounded by ϵ. This characterization not only expands our previous work [5] to the symmetric two source problem, but also characterizes the entire rate region in the non-asymptotic regime, not just coding with complete side information [5]. We show that the non-asymptotic rate region can be obtained by translating the asymptotic rate region away from the origin, as depicted in Figure 1.

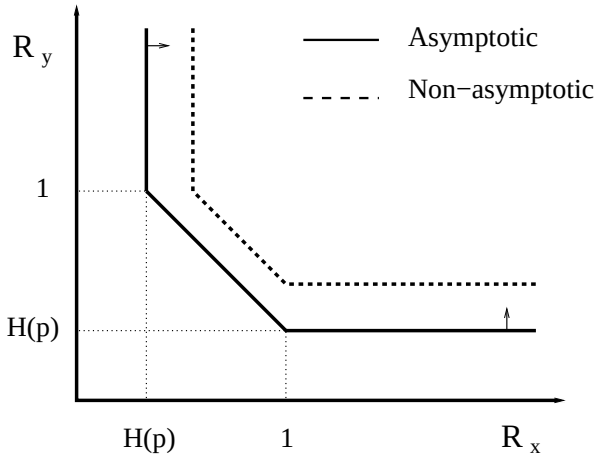

Figure 1: *The feasible rate regions for Slepian-Wolf coding for binary sequences with a symmetric distribution.*

We also consider the extension to the asymmetric two source problem. The problem of symmetric binary sources is relatively easy to study, because the joint probability can be characterized by a single sufficient statistic (details in Section IV). The asymmetric case poses more challenges, and we have started investigating this problem. Section V summarizes our progress in this direction.

The organization of this paper is as follows. We state the problem of non-asymptotic Slepian-Wolf coding in Section II. In Section III, we state the main result of our work, where we provide tight converse and achievable bounds. These bounds are proved in Section IV. We summarize our ongoing work in Section V and conclude in Section VI.

II. Problem statement

Consider two length-n binary sequences x and y that are correlated. The sequences x and y are both independent and identically distributed (iid) with Bernoulli(0.5). We study the case in which the correlation structure is also memoryless in the following sense: the sequence $z = x \oplus y$ is Bernoulli(p). We investigate the problem of distributed compression of the sequences x and y encoded at rates R_x and R_y, respectively.

The setup described above can be used to model the observations of a binary field F that takes on two equi-probable

[1]This work was supported by AFOSR, NSF, ONR, and the Texas Instruments Leadership University Program. Web: www.dsp.rice.edu.

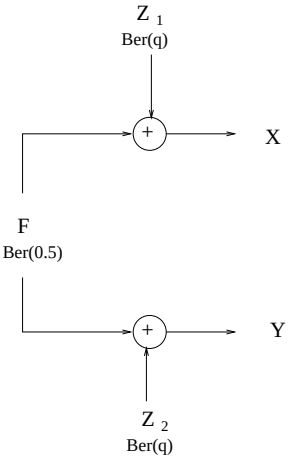

Figure 2: *Scenario 1: Sequences X and Y are noisy observations of the underlying field F. The Bernoulli parameter q is chosen such that $x \oplus y$ is Bernoulli(p).*

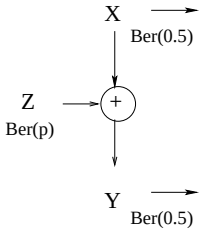

Figure 3: *Scenario 2: Sequence Y is obtained by superposition of X and noise Z.*

values 0 or 1. We make two noisy observations $x(i)$ and $y(i)$ of the field F as shown in Figure 2. The observations $x(i)$ and $y(i)$ have the same probability of error. We assume that the underlying field changes between every trial of the sequence, and is independent of the previous outcomes (this ensures iid x and y). We call this setup Scenario 1. Scenario 1 may occur in a simplified sensor network with two sensors that make noisy measurements over the same binary field. We are interested in compressing the data generated by the sensor network.

Another system that can be modeled by our setup is shown in Figure 3. We refer to this system as Scenario 2. The source X generates a length-n sequence x that is Bernoulli(0.5). The noise sequence z, which is Bernoulli(p), is added to x to create the sequence y.

Let the number of observations made in the sequence be n. The Law of Large Numbers states that for large n, the probability mass is concentrated close to the true distribution. By encoding only those sequences whose empirical distributions are close to the true distribution, we can make the probability of codeword error arbitrarily small. However, for finite n, we have non-zero probability of generating sequences whose type is substantially different from the true distribution. We consider two problems in the finite regime. The *converse* problem is to find conditions on (R_x, R_y) given that there is a Slepian-Wolf coding scheme with those rates. The *achievability* problem is to show that we can design a Slepian-Wolf coding scheme with rates R_x and R_y if (R_x, R_y) satisfies the conditions put forth by the converse. In Section III, we provide converse and achievable bounds that are tight, and therefore we completely characterize the non-asymptotic region for Slepian-Wolf coding for the symmetric distribution.

We define the rate R of a coding scheme as follows. For a Slepian-Wolf coding scheme that uses M bits to encode sequences of length n, we define the non-asymptotic rate as $R \triangleq M/n$. By providing tight converse and achievable bounds, we provide a feasible rate region (R_x, R_y) in which the probability of codeword error can be made smaller than ϵ.

Formally, the problem of finding the converse and achievable bounds is stated as:

Converse bound: Let there exist a Slepian-Wolf coding scheme for length-n sequences x and y that achieves probability ϵ of codeword error with rates R_x and R_y, respectively. Identify the conditions that R_x and R_y, and thus the feasible rate region, must satisfy.

Achievable bound: For a fixed n and any rate pair (R_x, R_y) that lies in the feasible rate region, construct a distributed source code with probability of codeword error bounded by ϵ.

III. Main result

The main result for the conditions on R_x and R_y for non-asymptotic Slepian-Wolf coding is given in the following theorem.

Theorem 1 *For length-n binary sequences x and y with a symmetric joint distribution, there exists a non-asymptotic Slepian-Wolf code with rates R_x and R_y if and only if the following conditions hold:*

$$R_x > H\left(p + \frac{\sqrt{p(1-p)}\phi^{-1}(\epsilon)}{\sqrt{n}}\right) + o(\frac{1}{\sqrt{n}}),$$

$$R_y > H\left(p + \frac{\sqrt{p(1-p)}\phi^{-1}(\epsilon)}{\sqrt{n}}\right) + o(\frac{1}{\sqrt{n}}),$$

$$R_x + R_y > 1 + H\left(p + \frac{\sqrt{p(1-p)}\phi^{-1}(\epsilon)}{\sqrt{n}}\right) + o(\frac{1}{\sqrt{n}}), \quad (1)$$

where ϵ is the probability of codeword error and p is the parameter of the Bernoulli sequence $x \oplus y$.[2]

The asymptotic rate region is given by $R_x \geq H(p)$, $R_y \geq H(p)$, and $R_x + R_y \geq 1 + H(p)$. Therefore, the non-asymptotic rate region can be obtained by translating the asymptotic rate region away from the origin. Both rate regions are illustrated in Figure 1. Note that, as n increases, the non-asymptotic rate region converges to the asymptotic region.

IV. Proof of bounds

IV.A Converse Bound

For the converse bound, we must show that the existence of a Slepian-Wolf code with rates R_x and R_y implies that the rates satisfy the conditions in (1). To prove this, consider Scenario 2, which is shown in Figure 3, with $z = x \oplus y$. If x and y have a symmetric joint distribution, then x and z are independent.

[2]For two functions $f(n)$ and $g(n)$, $f(n) = o(g(n))$ if for all positive $c > 0$, $\exists n_0 \in \mathbb{R}^+$, $0 \leq f(n) < cg(n)$ for all $n > n_0$. Similarly $f(n) = O(g(n))$ if $\exists c, n_0 \in \mathbb{R}^+$, $0 \leq f(n) \leq cg(n)$ for all $n > n_0$.

First consider the lower bound on R_x. In order to minimize the rate R_x, we consider lossless transmission of y. In our example, the minimum bit-rate R_y to achieve this is $H(0.5) = 1$, because y is Bernoulli(0.5). We can thus transmit y uncompressed: that is, we use y as complete side information. We have studied the problem of non-asymptotic Slepian-Wolf coding with complete side information [5]. Using our previous work, we conclude that $R_x > H\left(p + \frac{\sqrt{p(1-p)}\phi^{-1}(\epsilon)}{\sqrt{n}}\right) + o(\frac{1}{\sqrt{n}})$, which is the condition given in (1). By symmetry, the condition for R_y is also given by $R_y > H\left(p + \frac{\sqrt{p(1-p)}\phi^{-1}(\epsilon)}{\sqrt{n}}\right) + o(\frac{1}{\sqrt{n}})$. The intuition behind the above bounds is that the smallest set T_z such that $\Pr(z \in T_z) > 1 - \epsilon$ satisfies

$$\log(|T_z|) = n\left[H\left(p + \phi^{-1}(\epsilon)\sqrt{\frac{p(1-p)}{n}}\right) + o(\frac{1}{\sqrt{n}})\right].$$

To derive the lower bound on the sum rate $R_x + R_y$, consider the joint distribution of x and y. Define the minimal cardinality set $T_{\min}$ as the smallest set of pairs of sequences (x, y) that cover a probability of $1 - \epsilon$. Clearly, the sum rate $R_x + R_y$ is lower bound by $\frac{1}{n}\log|T_{\min}|$. The key observation we use to compute $|T_{\min}|$ is that the joint probability of (x, y) depends only on the sequence z. In particular, $|T_{\min}|$ depends only on the sufficient statistic n_z, where n_z is the number of 1's in the sequence z. This property holds because the joint distribution of x and y is symmetric: in other words the probability of observing $(x(i), y(i)) = (0, 0)$ is equal to the probability of observing $(x(i), y(i)) = (1, 1)$; similarly the probability of observing $(x(i), y(i)) = (0, 1)$ is equal to the probability of observing $(x(i), y(i)) = (1, 0)$. The pairs of sequences that have the largest probability are those that have $n_z = 0$, or equivalently, $x = y$. We accumulate the large probability pairs by increasing n_z until we cover a probability of $1 - \epsilon$. The minimal cardinality set is of the form

$$T_{\min} = \{(x, y) : n_z < qn\}$$

where q is some threshold. The set $T_{\min}$ contains all sequences x and y that do not differ in more than nq bits out of the total n bits.

Clearly, n_z is binomially distributed with parameters n and p. For large n, we invoke the central limit theorem (CLT) and approximate the distribution of n_z by a Gaussian with mean np and variance $np(1-p)$. Using the CLT approximation for n_z, the size of the minimal cardinality set is given by

$$|T_{\min}| = 2^{n + nH\left(p + \phi^{-1}(\epsilon)\sqrt{\frac{p(1-p)}{n}} + o(\frac{1}{\sqrt{n}})\right)}.$$

Therefore we have

$$\begin{aligned} R_x + R_y &\geq \frac{1}{n}\log|T_{\min}| \\ &= 1 + H\left(p + \phi^{-1}(\epsilon)\sqrt{\frac{p(1-p)}{n}} + o(\frac{1}{\sqrt{n}})\right). \end{aligned}$$

This proves the converse part of the theorem.

IV.B Achievable Bound

We now study achievability by designing a coding scheme for rates that lie in the feasible region given by (1). Let R_x and R_y satisfy the conditions in (1). We propose a scheme in which the probability of codeword error is upper-bound by ϵ.

The scheme we propose is based on random binning [6]. We partition the space of sequences that can be generated by source X into 2^{nR_x} bins and the space of sequences generated by Y into 2^{nR_y} bins. We assign every sequence x to one of the 2^{nR_x} bins independently using a function f_x according to a uniform distribution. Similarly, we assign every sequence y to one of the 2^{nR_y} bins using a function f_y according to a uniform distribution. We reveal these functions to both the encoder and the decoder.

Note that there are 2^n sequences each of x and y. Different sequences can be assigned to the same bin but perfect reconstruction is still guaranteed (with probability of error ϵ) as explained below.

The encoding procedure is as follows: The encoder for X sends the index of the bin to which the sequence x belongs. Similarly, the encoder for Y sends the index of the bin to which y belongs. Hence, the coding rate for source X is R_x, and the rate for source Y is R_y.

Before we proceed with the decoding procedure, we define the notion of *joint minimality* for two sequences. Two length-n sequences $\widetilde{x}$ and $\widetilde{y}$ are jointly minimal if the number of 1's in the sequence $\widetilde{x} \oplus \widetilde{y}$ is less than nq, where

$$q = p + (\phi^{-1}(\epsilon) + o(1))\sqrt{\frac{p(1-p)}{n}}. \tag{2}$$

In other words, $\widetilde{x}$ and $\widetilde{y}$ are jointly minimal if $(\widetilde{x}, \widetilde{y}) \in T_{\min}$.

The decoding procedure is as follows: Given a received index pair (i, j), we declare $(\widehat{x}, \widehat{y}) = (x, y)$ if there is a single pair of sequences (x, y) such that $f_x(x) = i$, $f_y(y) = j$ and the sequences x and y are jointly minimal. Otherwise, we declare an error.

To compute the probability of codeword error, we consider the possibilities that generate an error. Codeword error occurs in one of the following two cases:

- the sequences x and y are not jointly minimal, or
- there is another jointly minimal pair in the same pair of bins (i, j).

Let the probability that the sequences are not jointly minimal be ϵ_1. We have $\epsilon_1 = \phi(\phi^{-1}(\epsilon) + o(1)) + o(1) \approx \epsilon$. We choose the $o(1)$ term in (2) such that $\epsilon_1 < \epsilon$.

We now compute the probability of finding another minimal pair in the same pair of bins. The total number of jointly minimal pair of sequences is given by

$$|T_{\min}| = 2^{n\left(1 + H\left(p + \sqrt{p(1-p)}\phi^{-1}(\epsilon)/\sqrt{n} + o(\frac{1}{\sqrt{n}})\right)\right)}. \tag{3}$$

Let the probability that there is another jointly minimal pair of sequences in the same pair of bins be p_{conflict}. To determine p_{conflict}, we note that there is no conflict if all the other jointly minimal sequences occupy other bin pairs. The total number of pairs of bins is given by $N_{\text{bins}} = 2^{n(R_x+R_y)}$. The probability that another jointly minimal pair of sequences lies in another bin pair is given by $1 - 2^{-n(R_x+R_y)}$. Consequently,

$$1 - p_{\text{conflict}} = \left(1 - 2^{-n(R_x+R_y)}\right)^{|T_{\min}|}.$$

As long as $|T_{\min}|2^{-n(R_x+R_y)} \ll 1$, the quantity p_{conflict} is small. The condition $R_x + R_y > \log(|T_{\min}|) + O(\frac{1}{n})$ suffices; the $o(\frac{1}{\sqrt{n}})$ in (1) includes the region $R_x + R_y > \log(|T_{\min}|) + O(\frac{1}{n})$. In this regime, the probability of finding another jointly minimal pair in the same pair of bins can be made as small as necessary by adding $O(\frac{1}{n})$ to the sum rate.

The total probability of codeword error is $\epsilon_1 + p_{\text{conflict}}$. By choosing the $o(\cdot)$ terms appropriately, we have a coding scheme that achieves probability of error upper bound by ϵ. This completes the proof of Theorem 1. □

V. Beyond symmetric distributions

The main result of this paper as stated in Theorem 1 applies to two symmetric binary sources. In order to prove the converse and achievable bounds in Theorem 1, we used the crucial property that the joint probability of sequences x and y depends only on the number of 1's in the sequence $z = x \oplus y$. In this section, we provide directions that we are pursuing to derive the Slepian-Wolf bounds for two asymmetric sources.

For a asymmetric pair of sequences, we have the probabilities of the outcomes given by $\Pr\{(x(i), y(i)) = (0,0)\} = q_1$, $\Pr\{(x(i), y(i)) = (0,1)\} = q_2$, $\Pr\{(x(i), y(i)) = (1,0)\} = q_3$, and $\Pr\{(x(i), y(i)) = (1,1)\} = q_4$. Note that for the symmetric case, we had $q_1 = q_4$ and $q_2 = q_3 (= p)$. To derive the bounds for the general case, we consider two sets in the probability distribution: the typical set T_{typ} and the minimal cardinality set $T_{\min}$. Given ϵ, we define the typical set T_{typ} as the set of all pairs of sequences that are close to the true distribution (in terms of divergence) that covers a probability $1 - \epsilon$. The minimal cardinality set $T_{\min}$ is the smallest set of pairs of sequences that covers a probability $1 - \epsilon$.

Although both sets T_{typ} and $T_{\min}$ cover a probability of $1-\epsilon$, let us highlight the important difference. The typical set T_{typ} contains pairs of sequences (x, y) that have approximately the same probability of occurrence, close to $2^{-nH(X,Y)}$. On the other hand, the pairs of sequences in the minimal cardinality set $T_{\min}$ have widely different probabilities of occurrence; in particular, $T_{\min}$ contains the most likely pairs including those for which $x = y$. The set $T_{\min}$ is the smallest set among all sets of joint sequences that cover a probability $1 - \epsilon$. In particular, $|T_{\min}| \leq |T_{\text{typ}}|$.

Given a rate R, the best coding strategy that minimizes the probability of codeword error is to encode only the sequences in the minimal cardinality set $T_{\min}$. Clearly, the sum rate $R = R_x + R_y$ cannot be smaller than $\frac{\log(|T_{\min}|)}{n}$.

It can be shown that the typical set T_{typ} can be described by types that are bounded by an ellipse centered around the true distribution Q. For a typical set corresponding to a given ϵ, we can easily compute the cardinality $|T_{\text{typ}}|$. We use this result to derive the cardinality of the set $T_{\min}$ indirectly, because the direct derivation of $|T_{\min}|$ appears to be quite involved.

In order to characterize $T_{\min}$, consider a multinomial distribution of alphabet size 4 with n trials. The true distribution is given by the vector $Q = [q_1, q_2, q_3, q_4]$. The key is to find the minimal set that covers probability $1 - \epsilon$, where ϵ is the probability of codeword error. For large n, the multinomial distribution tends to a Gaussian (that is, various empirical statistics tend to have a Gaussian distribution), owing to the Central Limit Theorem. To compute $|T_{\min}|$, we calculate the portion of the Gaussian distribution that needs to be retained and the remaining portion that needs to be truncated.

The set $T_{\min}$ consists of the sequences with the largest probabilities. The probability of a sequence is given by $2^{-n(H+D)}$, where H is the sample entropy of the type, and D is the divergence from the center of the 3-dimensional ellipse. The set $T_{\min}$ is therefore the set of all points such that $H + D < \tau$, where τ is a constant. The minimal cardinality set is thus bounded by a hyperplane, given by $H + D = \tau$. Note that the set $H + D = \tau$ is a hyperplane because $H + D$ is the coding length given actual probability Q and empirical Q', and the value of $H + D$ is linear in Q'. These insights are summarized in Figure 4. Our sketch is 2-dimensional for ease of visualization.

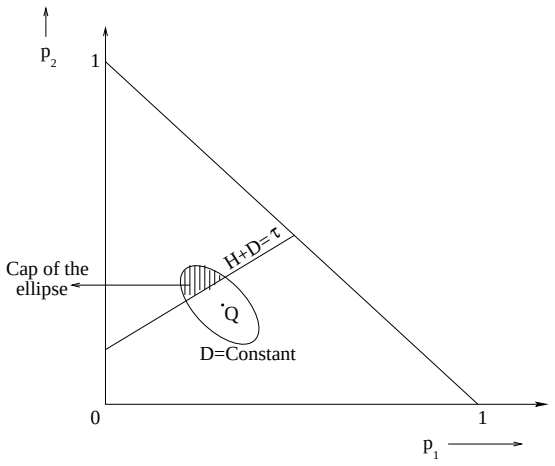

Figure 4: *Consider for illustration a 2-dimensional probability simplex given by the triangle. The ellipse shown has constant divergence D about the true distribution Q and covers a probability $1 - \epsilon'$. The minimal cardinality set is bounded by the hyperplane $H + D = \tau$. The cap of the ellipse (shaded) is the region in the simplex that lies within the ellipse but outside the minimal cardinality set. We are interested in computing the probability of the cap of the ellipse.*

In order to compute the probability of codeword error, we consider the typical set T_{typ} and the minimal cardinality set $T_{\min}$. As long as T_{typ} covers the vast majority of the probability space, the main error term is the portion of the ellipse T_{typ} that lies outside $T_{\min}$. The computation of the probability of this event requires integration over the cap of the ellipse. Our current work is focused on deriving analytical and numerical solutions for this problem.

VI. Summary and conclusions

Our contribution in this work is the determination of achievable and converse bounds for Slepian-Wolf coding for two symmetric binary sources. In this setup, we completely characterized the rate region (R_x, R_y) for non-asymptotic Slepian-Wolf coding in which the probability of codeword error is bounded by ϵ. We showed that the non-asymptotic rate region can be obtained by translating the asymptotic rate region away from the origin. We also provided directions on how this work can be extended to handle asymmetric binary sources.

Acknowledgments

Thanks to Zixiang Xiong of Texas A&M University, who suggested the symmetric problem, which is of great interest because it is widely used to model communication sources.

References

[1] S. S. Pradhan and K. Ramchandran, "Distributed source coding: Symmetric rates and applications to sensor networks," *Proc. IEEE Data Compression Conference* (DCC), pp. 363–372, Mar. 2000.

[2] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Mag.*, vol. 21, pp. 80–94, Sept. 2004.

[3] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Trans. Information Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

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

[5] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "Redundancy rates of Slepian-Wolf coding," *Proc. 42nd Annual Allerton Conference on Communication, Control, and Computing*, Oct. 2004.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, New York, 1991.