Redundancy Rates of Slepian-Wolf Coding*

Dror Baron, Mohammad Ali Khojastepour, and Richard G. Baraniuk
Dept. of Electrical and Computer Engineering, Rice University, Houston, TX 77005

Abstract

The results of Shannon theory are asymptotic and do not reflect the finite codeword lengths used in practice. Recent progress in design of distributed source codes, along with the need for distributed compression in sensor networks, has made it crucial to understand how quickly distributed communication systems can approach their ultimate limits. Slepian and Wolf considered the distributed encoding of length-n sequences x and y. If y is available at the decoder, then as n increases x can be encoded losslessly at rates arbitrarily close to the conditional entropy $H(X|Y)$. However, for any finite n there is a positive probability that x and y are not jointly typical, and so x cannot be decoded correctly. We examine a Bernoulli setup, where x is generated by passing y through a binary symmetric correlation channel. We prove that the finite n requires us to increase the rate above the conditional entropy by $K(\epsilon)/\sqrt{n}$, where ϵ is the probability of error. We also study the cost of universality in Slepian-Wolf coding, and propose a universal variable rate scheme wherein the encoder for x receives $P_Y = \frac{1}{n}\sum_i y_i$. For $P_Y < 0.5$, our redundancy rate is $K'(\epsilon)/\sqrt{n}$ above the empirical conditional entropy. When $|P_Y - 0.5| = O(n^{-1/6})$, $K'(\epsilon) = \Omega(n^{1/6})$, and another scheme with redundancy rate $O(n^{-1/3})$ should be used. Our results indicate that the penalties for finite n and unknown statistics can be large, especially for $P_Y \approx 0.5$.

1 Introduction

1.1 Motivation

Distributed source coding [1–5] relies on the availability of side information at the decoder. In the Slepian-Wolf framework [1–3, 6], this side information enables the encoder to communicate losslessly at the conditional entropy rate, rather than the entropy. For example, sensor networks [4, 5] rely on data that often exhibit strong spatial correlations, and can thus benefit greatly from distributed compression. Furthermore, as individual sensors are often battery-operated, low power consumption is a limiting factor, and the reduction of communication costs via distributed compression is important.

Shannon theory [1–3] has provided a variety of bounds on the efficiency of communication systems. Unfortunately, most of these results rely on asymptotics; the bounds reflect the limiting performance using infinitely long codewords. In contrast, practical communication systems live in a non-asymptotic world; they must use finite computational resources, incur finite delay, and so on. Therefore, in order to use these resources most efficiently, it has become crucial to understand how quickly practical communication systems can approach their ultimate performance limits in the non-asymptotic regime.

Another problem that is sometimes overlooked in Shannon theory is lack of availability of the source and channel statistics. In lossless source coding [1, 3], an inaccurate model

*This research was supported by AFOSR, ONR, NSF, and the Texas Instruments Leadership University Program

only incurs a coding length penalty. In contrast, channel coding and distributed coding rely on joint typicality; the use of a model that does not "cover" the true statistics may result in an error. Consequently, many practical systems use models that are perhaps too robust to model mismatch, resulting in poor performance. Therefore, universal methods whose rates converge to the asymptotic performance limits are highly desirable.

1.2 Non-asymptotic Slepian-Wolf rate

Practical codes use codewords of a finite length n (n may be large [4, 7]), and the probability of error (either the bit error rate or codeword error rate ϵ) must be reasonably small. With these conditions on n and ϵ, it is not clear what rates are actually achievable, because Shannon theory has not addressed this point. Despite being an imprecise measure for practical channel [7] and distributed [4, 5] codes, the capacity and distributed coding rates are the standard benchmarks for verifying the efficiency of coding techniques.

Consider the question: "If the codeword length n is given, then what is the minimum possible rate R of coding x with side information y at the decoder, for which the probability of codeword error is bounded by a given ϵ?" We denote this *non-asymptotic Slepian-Wolf rate* by $R_{\mathrm{NA}}(n, \epsilon)$. This definition provides a more informative benchmark for verifying the efficiency of practical distributed codes.

1.3 Related results

Although the non-asymptotic rate has not been clearly stated in the literature, there exist several related results. To the best of our knowledge, the converse and achievable proofs for Slepian-Wolf coding by Wolfowitz [3] are the only directly related bounds in the literature. However, Wolfowitz did not identify the notion of non-asymptotic rate, and may not have realized its importance. Furthermore, his bounds are loose. In contrast, our bounds are precise to within $o(1/\sqrt{n})$ rate.[1]

Error exponent bounds [8, 9], which provide lower and upper bounds on the probability ϵ of block error, can be viewed as dual to our results. However, for rates near the asymptotic performance limits, the lower and upper bounds provided by error exponents are often relatively loose (details in Baron et al. [10]). Our results, which are based on the central limit theorem (CLT), are more effective for rates near (within $O(1/\sqrt{n}$ of the) Slepian-Wolf rate. Our results are less effective for rates substantially above this rate. In fact, for such rates ϵ becomes very small, and the CLT no longer applies.

1.4 Contributions

We focus on a setup wherein a Bernoulli sequence x is generated by passing side information y through a binary symmetric correlation channel z, which is independent of y. This setup should be viewed as an illustration of a broader concept.

In Section 3 we prove that $R_{\mathrm{NA}}(n, \epsilon) = H(X|Y) + [K(\epsilon) + o(1)]/\sqrt{n}$. In Section 4 we consider the penalty for unknown statistics. We construct a universal Slepian-Wolf coding scheme whose variable rate is based on the empirical statistic $P_Y = \frac{1}{n}\sum_{i=1}^{n} y_i$. For P_Y that is bounded away from 0.5, the redundancy of this scheme above the empirical conditional entropy is comparable to the redundancy with known source statistics. However, when

[1] For two functions $f(n)$ and $g(n)$, $f(n) = o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$. Similarly, $f(n) = O(g(n))$ if $\exists C \in \mathbb{R}^+,\ \lim_{n\to\infty} |f(n)/g(n)| \leq C$, and $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$.

$|P_Y - 0.5| = O(n^{-1/6})$, the redundancy rate becomes $\Omega(n^{-1/3})$, and another scheme with redundancy rate $O(n^{-1/3})$ should be used. Therefore, the penalty for universality may be quite large in practice.

Our results do not contradict the traditional Slepian-Wolf analysis [1, 2, 6]. Rather, they refine and strengthen the classical results, because $\lim_{n\to\infty} R_{\mathrm{NA}}(n, \epsilon) = H(X|Y)$ for any $\epsilon \in (0, 1)$. Whereas error exponents [8, 9] characterize the block error probability well for rates substantially above the Slepian-Wolf rate, our bounds are more precise for rates near that limit. Finally, the relations between this paper and the analysis of the non-asymptotic capacity of the binary symmetric channel [10] suggest that the redundancy of universal channel coding systems with feedback could turn out to be exceedingly high.

We begin by defining our problem setup formally in Section 2. We then provide results on the non-asymptotic rate with known source statistics in Section 3. Section 4 evaluates our universal Slepian-Wolf scheme.

2 Coding with side information

Let $x = x_1, \ldots, x_n$, $y = y_1, \ldots, y_n$, and $z = z_1, \ldots, z_n$ be length-n *Bernoulli sequences*, which are independent and identically distributed over the binary alphabet $\{0, 1\}$. These sequences satisfy $x_i = y_i \oplus z_i$, where $\oplus$ denotes modulo-2 addition, and the sequence z is independent of y. We thus interpret z as a binary symmetric *correlation channel* from y to x. Let $p = \Pr(y_i = 1)$, $q = \Pr(z_i = 1)$, and $r = \Pr(x_i = 1)$ be the *Bernoulli parameters*, where $p, q \in (0, 0.5)$ and $r = p(1 - q) + (1 - p)q$. Let X, Y, and Z denote random variables for the sequences x, y, and z, respectively. The per-symbol *joint entropy* [1] of X and Y satisfies $H(X, Y) = H(Y) + H(X|Y) = H(Y) + H(Z)$, where $H(Z) = H_2(p)$ and $H_2(p) = -p \log(p) - (1 - p) \log(1 - p)$ is the *binary entropy*. In order to encode both x and y, the expected per-symbol coding length must exceed $H(X, Y)$. Furthermore, it can be shown that $H(X) > H(Z)$ when y is not deterministic, i.e., when $H(Y) > 0$.

Now consider a coding system in which the *encoder* for x sends the index $f(x) \in \{1, \ldots, M\}$ to the decoder, where the *side information* y is available at the decoder but not at the encoder. The value of M is the *codebook size*, and $R_x = \log(M)/n$ is the *rate* required to encode x. The lossless encoding of the side information y can be performed using a variety of well-known techniques [1]; we thus concentrate on the encoding of x in the sequel. The *decoder* receives y and $f(x)$, and attempts to reconstruct x by $\widehat{x} \in \{0, 1\}^n$ via a mapping $\widehat{x} = g(y, f(x))$.

Slepian and Wolf [2] proved that asymptotically, as the block n increases, it suffices for R_x to exceed the *conditional entropy* $H(X|Y) = H(X, Y) - H(Y) = H(Z) = H_2(p)$ [1], which is less than $H(X)$ for $H(Y) > 0$. Cover [1, 6] used random binning to prove the result of Slepian and Wolf [2]. In the encoder, each input is randomly assigned an index $f(x)$ in a set $\{1, \ldots, M\}$, where $M = 2^{nR_x}$, $R_x = H(X|Y) + \delta$, and $\delta > 0$. In the decoder, if there is a single length-n sequence $\widehat{x}$ that is *jointly typical* with y and is assigned the same index, then the decoder reconstructs x by $\widehat{x}$. If there are multiple or no such sequences, then the decoder declares an error. Cover's proof [6] that the probability of error is small relies on the observation that it is unlikely that there are additional sequences that are assigned the same index and are jointly typical with y.

Unfortunately, for any finite n there is a positive probability that z is atypical, in which case x is not jointly typical with y. Therefore, the probability of error ϵ is strictly positive. We denote a Slepian-Wolf encoder and decoder pair (f, g) that uses codewords of length n, codebook size M, and achieves a probability of error no larger than ϵ by

(f, g, n, M, ϵ). For finite n, we define the *non-asymptotic rate* as

$$R_{\mathrm{NA}}(n, \epsilon) \triangleq \min_{\left\{\exists \text{ code } (f, g, n, \widetilde{M}, \epsilon)\right\}} \log(\widetilde{M})/n.$$

This definition provides a more informative benchmark for verifying the efficiency of practical Slepian-Wolf codes. Our objective is to provide bounds on $R_{\mathrm{NA}}(n, \epsilon)$.

3 Coding theorems

In this section we prove that $R_{\mathrm{NA}}(n, \epsilon) = H_2(q) + K(\epsilon)/\sqrt{n} + o(1/\sqrt{n})$, where $K(\epsilon)$ is specified in closed form. This provides a relatively precise characterization of the non-asymptotic Slepian-Wolf rate. Our results extend and refine the work of Wolfowitz [3], who proved that $R_{\mathrm{NA}}(n, \epsilon) > H(X|Y) + K_{\mathrm{c}}(\epsilon)/\sqrt{n}$ and $R_{\mathrm{NA}}(n, \epsilon) < H(X|Y) + K_{\mathrm{a}}(\epsilon)/\sqrt{n}$. Because $K_{\mathrm{a}}(\epsilon) > K_{\mathrm{c}}(\epsilon) > 0$, the *converse* and *achievable* bounds of Wolfowitz are loose. In contrast, we provide the same $K(\epsilon)$ for both converse and achievable bounds.

3.1 Mathematical preliminaries

Before providing our converse and achievable bounds, we characterize the typical set of values for the correlation channel sequence z. Let $n_z \triangleq \sum_{i=1}^{n} z_i$. Because z is Bernoulli, n_z has a binomial distribution, which can be approximated by a Gaussian distribution with mean nq and variance $nq(1-q)$. Consider the value t that n_z attains $\Phi^{-1}(\epsilon) + o(1)$ standard deviations above the mean, where Φ^{-1} is the inverse Gaussian error function and the $o(1)$ term decays to zero as n increases. We have

$$t \triangleq nq + [\Phi^{-1}(\epsilon) + o(1)]\sqrt{nq(1-q)}, \tag{1}$$

where the significance of the $o(1)$ term will be explained shortly. Define the typical set $T \triangleq \{z : \ n_z \leq t\}$; the atypical set is $T^C \triangleq \{z : \ n_z > t\}$. The CLT asserts that

$$\Pr(z \in T^C) = \Phi(\Phi^{-1}(\epsilon) + o(1)) + o(1) \approx \epsilon, \tag{2}$$

where the first $o(1)$ term is chosen such that $\Pr(z \in T^C)$ is slightly larger or smaller than ϵ, as required in the sequel.

Because $\Pr(z)$ is monotone decreasing in n_z, any set T' such that $\Pr(z \in T') \geq \Pr(z \in T)$ has cardinality $|T'|$ no smaller than $|T|$. Therefore, in order to attain a probability of error no larger than ϵ, the encoder must effectively assign different indices to at least $|T|$ possible inputs. The following Lemma describes the cardinality of T.

Lemma 1 [3, 10] *The cardinality of the typical set T satisfies*

$$\log(|T|) = nH(t/n) - 0.5\log(n) + O(1).$$

3.2 Non-asymptotic rate of coding with side information

We now state the main result of this section, which is proved in Appendix A. Note that Theorem 1 also applies to distributions on y that are not Bernoulli.

Theorem 1 *For a binary symmetric correlation channel with cross-over probability* $q \in (0, 0.5)$*, the non-asymptotic rate of a Slepian-Wolf code satisfies*

$$R_{\mathrm{NA}}(n, \epsilon) = H\left(q + \Phi^{-1}(\epsilon)\sqrt{\frac{q(1-q)}{n}}\right) + o(1/\sqrt{n}). \tag{3}$$

Our proof has converse and direct parts. The **converse part** shows that the existence of a Slepian-Wolf code (f, g, n, M, ϵ) implies the existence of a *fixed rate source code* for z that achieves a probability of error no larger than ϵ. We then lower bound by $|T|$ the codebook size M of a source code for z that achieves a probability of error no larger than ϵ. Therefore, the non-asymptotic rate is lower-bounded by $\log(|T|)/n$, and the converse part of the proof is completed by invoking Lemma 1.

The **direct part** of the proof constructs a Slepian-Wolf code that achieves a probability of error smaller than ϵ using a codebook size $M = 2^{nH(t/n)+o(\sqrt{n})}$. The encoder $f(x)$ assigns x an index randomly. The decoder searches for the single input $\widehat{x}$ that satisfies $\widehat{z} = \widehat{x} \oplus y \in T$ and $f(x') = f(x)$. We complete the proof by analyzing the two possible causes of error. In the first case, $z \notin T$; by choosing the $o(1)$ term in (1) appropriately, this happens with probability smaller than ϵ. In the second case, there exist additional inputs that satisfy the above decoding conditions; by choosing the $o(\sqrt{n})$ term in M to be "large enough", the probability of this event becomes negligible.

3.3 Redundancy rate of coding with side information

We define the *redundancy rate* $\mathcal{R}(n, \epsilon)$ as the excess rate of coding with side information in the non-asymptotic regime above the asymptotic Slepian-Wolf rate, i.e., $\mathcal{R}(n, \epsilon) \triangleq R_{\mathrm{NA}}(n, \epsilon) - H(X|Y)$. Theorem 1 indicates that $R_{\mathrm{NA}}(n, \epsilon) = H(X|Y) + K(\epsilon)/\sqrt{n} + o(1/\sqrt{n})$. Therefore, $\mathcal{R}(n, \epsilon) = K(\epsilon)/\sqrt{n} + o(1/\sqrt{n})$. The detailed results are provided in closed form in Theorem 2. The derivation of $K(\epsilon)$, which also appears in Baron et al. [10], relies on a first-order Taylor approximation of the binary entropy around q.

Theorem 2 [10] *For a binary symmetric correlation channel with cross-over probability* $q \in (0, 0.5)$*, the redundancy rate of coding with side information satisfies*

$$\mathcal{R}(n, \epsilon) \;=\; \frac{K(\epsilon)}{\sqrt{n}} + o(1/\sqrt{n}),$$

where the constant $K(\epsilon)$ *is given by*

$$K(\epsilon) = \Phi^{-1}(\epsilon)\sqrt{q(1-q)}\log\left(\frac{1-q}{q}\right).$$

For $\epsilon < 0.5$ the non-asymptotic Slepian-Wolf rate is greater than the conditional entropy, and the redundancy rate is positive and strictly proportional to $1/\sqrt{n}$.

4 Universality

4.1 Compression framework

We now move to the setting in which the statistics for x, y, and z are unknown. Our only assumptions are that these sequences are Bernoulli, that y and z are independent,

and that $x = y \oplus z$. Our goal is to derive codes that have reasonably low redundancies for a range of possible source statistics.

When the source statistics are unknown, the code must be robust to a range of possible statistics, including high-entropy statistics for which $H(X|Y)$ is large. Unfortunately, such statistics require high rates. In Section 3.2 we used fixed rate codes for coding x with side information y. A universal fixed rate code must use a high rate, hence the redundancy for low-entropy sources is large. Instead, we use *variable rate codes*, wherein the coding rate is adapted to the conditional entropies of the sources.

Oohama [9] (see also Kimura and Uyematsu [11]) described a Slepian-Wolf coding system with linked encoders. In this setup, in addition to the encoders described in Sections 2 and 3, there are additional encoders that transfer information between the main encoders for x and y. These encoders use low rates and can help the main encoders for x and y estimate the joint statistics.

Our setup is based on linked encoders. The encoder for the side information y communicates n_y, the number of 1's in y, to the encoder for x. Because $n_y \in \{0, 1, \ldots, n\}$, the description of n_y requires approximately $\log(n)$ bits, which is insignificant relative to other redundancy terms. We now describe the encoding of x, which is the main challenge.

4.2 Encoder for x

The direct part of the proof of Theorem 1 showed that the code construction is straightforward once the codebook size M is known. Furthermore, the selection of M relies on the distribution of n_z. However, the encoder for x has access to n_y and x, whereas the Bernoulli parameter q of the correlation channel sequence z is unknown. Therefore, the gist of the problem is to estimate q and the distribution of n_z from x and n_y.

The encoder can use n_y and x to estimate the distribution of n_z. Because x and y are Bernoulli, pairs of sequences with the same values of n_x, n_y, and n_z have the same probability. But n_z is unknown, so the encoder cannot distinguish between different pairs of sequences with the same n_x and n_y. Therefore, the distribution of n_z, and hence the size of the codebook, is based solely on n_x and n_y. We thus define the encoder for x as a mapping $f(x, n_y) \in \{1, \ldots, M_{n_x,n_y}\}$, where the codebook size M_{n_x,n_y} varies with n_x and n_y. We now focus on the distribution of n_z, which is needed to determine M_{n_x,n_y}.

4.3 Distribution of n_z

The conditional probability $\Pr(n_z|n_x, n_y)$ can be simplified as

$$\begin{aligned}\Pr(n_z|n_x, n_y) &= \frac{\Pr(n_x, n_y, n_z)}{\sum_k \Pr(n_x, n_y, k)} = \frac{\Pr(n_x|n_y, n_z)\Pr(n_y)w(n_z)}{\sum_k \Pr(n_x|n_y, k)\Pr(n_y)w(k)} \\ &= \frac{w(n_z)\Pr(n_x|n_y, n_z)}{\sum_k w(k)\Pr(n_x|n_y, k)}, \end{aligned} \tag{4}$$

where we use priors $\Pr(n_y)$ and $w(n_z)$ for n_y and n_z.

In order to evaluate the conditional probability $\Pr(n_x|n_y, n_z)$, we first define $P_X \triangleq n_x/n$, $P_{X'} \triangleq 1 - P_X$, $P_Y \triangleq n_y/n$, $P_{Y'} \triangleq 1 - P_Y$, $P_Z \triangleq n_z/n$, and $P_{Z'} \triangleq 1 - P_Z$. These are the *empirical* first order statistics of the data. Similarly, we define the empirical joint (second order) statistics $P_{Y',Z'} \triangleq \frac{1}{n}\sum_{i=1}^n 1_{\{y_i=0,z_i=0\}}$, $P_{Y',Z} \triangleq \frac{1}{n}\sum_{i=1}^n 1_{\{y_i=0,z_i=1\}}$, $P_{Y,Z'} \triangleq \frac{1}{n}\sum_{i=1}^n 1_{\{y_i=1,z_i=0\}}$, $P_{Y,Z} \triangleq \frac{1}{n}\sum_{i=1}^n 1_{\{y_i=1,z_i=1\}}$ and the empirical conditional statistics $P_{Z'|Y'} \triangleq P_{Z',Y'}/P_{Y'}$, $P_{Z|Y'} \triangleq P_{Z,Y'}/P_{Y'}$, $P_{Z'|Y} \triangleq P_{Z',Y}/P_Y$, and $P_{Z|Y} \triangleq P_{Z,Y}/P_Y$. The

encoder for x knows P_X, $P_{X'}$, P_Y, and $P_{Y'}$, but does not know P_Z, $P_{Z'}$, or any of the joint or conditional empirical statistics. Nonetheless, if P_Z were known, then the empirical joint and conditional statistics can easily be computed. Using a combinatorial argument,

$$\Pr(n_x|n_y,n_z) = \frac{\binom{nP_Y}{nP_{Y,Z}}\binom{nP_{Y'}}{nP_{Y',Z}}}{\binom{n}{nP_Z}}.$$

A derivation based on Stirling's formula $k! = \sqrt{2\pi k}(k/e)^k[1+O(1/k)]$ then provides

$$\binom{a}{b} = \frac{2^{nH_2(b/a)}}{\sqrt{2\pi a P_B P_{B'}}}\left[1+O\left(\frac{1}{aP_BP_{B'}}\right)\right],$$

where $P_B = b/a$ and $P_{B'} = (a-b)/a$. We now have

$$\begin{aligned}\Pr(n_x|n_y,n_z) &= \frac{2^{nP_YH_2(P_{Z|Y})}2^{nP_{Y'}H_2(P_{Z|Y'})}\sqrt{2\pi nP_ZP_{Z'}}}{2\pi n\sqrt{P_YP_{Z|Y}P_{Z'|Y}P_{Y'}P_{Z|Y'}P_{Z'|Y'}}2^{nH_2(P_Z)}}\left[1+O\left(\frac{1}{n}\right)\right] \quad (5)\\ &= 2^{-nI(P_Z;P_Y)}\sqrt{\frac{P_ZP_{Z'}}{2\pi nP_YP_{Y'}P_{Z|Y}P_{Z'|Y}P_{Z|Y'}P_{Z'|Y'})}}\left[1+O\left(\frac{1}{n}\right)\right] \quad (6)\end{aligned}$$

where in (5) we assume that the various empirical probabilities are bounded away from 0, and in (6) we use $I(P_Z;P_Y)$ to denote the mutual information [1] between the empirical joint statistics of Y and Z.

Let us pause and reflect on these expressions. When n is large, the second term in (6) varies slowly and the order term becomes insignificant. In contrast, the mutual information has a drastic effect on $\Pr(n_x|n_y,n_z)$. Therefore, this expression is maximal when $I(P_Z;P_Y)=0$, i.e., when y and z are empirically independent. We conclude that $P(n_x|n_y,n_z)$ is maximal when $P_{Z|Y} \approx P_{Z|Y'} \approx P_Z$. When this occurs, $P_X \approx P_YP_{Z'} + P_{Y'}P_Z = P_Y + P_Z(1-2P_Y)$, which implies that the maximum occurs when P_Z is close to

$$P_Z^* \triangleq \frac{P_X - P_Y}{1-2P_Y}.$$

In terms of $\Pr(n_z|n_x,n_y)$, it is reasonable to assume that $w(z)$ in (4) varies slowly in k. Therefore, $\Pr(n_z|n_x,n_y)$ peaks around P_Z^*, and the rate of decay near the peak is dominated by $2^{-nI(P_Z;P_Y)}$. A derivation based on the Fisher information [1] reveals that

$$I(P_Z^*+\Delta;P_Y) = \frac{\Delta^2(1-2P_Y)^2}{8\ln(2)P_YP_{Y'}P_Z^*P_{Z'}^*} + O(\Delta^3).$$

We conclude that as n increases, $\Pr(n_z|n_x,n_y)$ is similar to Gaussian with mean P_Z^* and variance $\frac{4nP_YP_{Y'}P_Z^*P_{Z'}^*}{(1-2P_Y)^2}$. Using the direct part of Theorem 1, we have the following result.

Theorem 3 *For a Bernoulli source y, a binary symmetric correlation channel z that is independent of z, a channel output x, and a uniform prior $w(n_z) = 1/n$, the non-asymptotic rate of our universal Slepian-Wolf code satisfies*

$$R_{\mathrm{NA}}(n,\epsilon) = H\left(P_Z^* + \Phi^{-1}(\epsilon)\sqrt{\frac{4P_YP_{Y'}P_Z^*P_{Z'}^*}{n(1-2P_Y)^2}}\right) + o(1/\sqrt{n}).$$

4.4 Discussion

Theorem 3 indicates that the redundancy of the universal algorithm is larger than the redundancy of the fixed rate scheme that relies on known statistics *by a factor* of $f(P_Y) = 2\sqrt{P_Y P_{Y'}}/|1-2P_Y|$. This factor provides two interesting twists. First, for small P_Y we have $f(P_Y) < 1$. In this regime, P_X provides a relatively precise prediction of P_Z, hence a low-redundancy source code can be constructed.[2] In contrast, a fixed rate code always estimates n_z with a variance $nq(1-q)$, even if $n_y = 0$. This result suggests that a variable rate Slepian-Wolf scheme based on a minor amount of side information ($\log(n)$ bits to describe n_y) may have substantially lower redundancy than a fixed rate scheme. In fact, it can be shown that incorporating an appropriate prior $w(n_z)$ into our universal method can reduce the expected redundancy when the source statistics are known.

The second surprising result is that when P_Y is close to 0.5, the redundancy of the universal scheme is much greater than that of the fixed rate code that relies on knowledge of the statistics. Actually, for $P_Y = 0.5 - \alpha$ we have $f(P_Y) = O(\alpha)$, hence the redundancy may be quite large. Consider the following numerical example for $n = 10^4$ and a correlation channel with parameter $q = 0.1$. The ideal Slepian-Wolf rate is $nH_2(q) = 4690$ bits. The fixed rate code of Section 3.2 requires $nR_{\mathrm{NA}}(n,\epsilon) = 4907$ bits for $\epsilon = 10^{-2}$. Now consider the universal code, where the empirical statistics are $P_Y = 0.3$ and $P_X = P_Y(1-q) + P_{Y'}q$. In this case the universal code needs 5224 bits, not including $\log(n) \approx 13$ bits for conveying n_y to the encoder. Finally, when $P_Y = 0.4$, the performance of the universal code deteriorates to 5863 bits. The redundancy now comprises 25% of the theoretical Slepian-Wolf coding length, although these values for n, ϵ, P_Z, and P_Y seem quite ordinary.

The fundamental problem in the regime in which P_Y is close to 0.5 is that it becomes increasingly difficult to estimate the statistics of the correlation channel. For fixed $P_Y \in (0, 0.5)$, as n increases the redundancy rate is $O(1/\sqrt{n})$. However, for finite n we may have $P_Y \approx 0.5$, and our estimate of P_Z is exceedingly noisy. In fact, when $|P_Y - 0.5| = O(n^{-1/6})$, $K'(\epsilon) = \Omega(n^{1/6})$, and the redundancy rate becomes $\Omega(n^{-1/3})$. In this situation we can use an alternative universal scheme that transmits the first k bits of y to the encoder as a training sequence, and then estimates P_Z to within a resolution of $O(1/\sqrt{k})$. This scheme provides a better estimate if k is large enough. We can provide a maximal universal redundancy rate $O(n^{-1/3})$ using this scheme.

Finally, the universal Slepian-Wolf setup with linked encoders is related to channel coding with feedback. Our results indicate that the estimation of channel statistics is exceedingly difficult when the channel input probability P_Y is close to 0.5. Unfortunately, the capacity-achieving channel codes for the binary symmetric channel use precisely this probability [1, 3, 8, 10]. Therefore, the redundancy rates of universal channel coding systems with feedback could turn out to be exceedingly high. In fact, when the source statistics are known, we must back off from the capacity by $O(1/\sqrt{n})$ [10]. In contrast, for $P_Y = 0.5$ the redundancy rate may be as large as $O(n^{-1/3})$. We leave the resolution of these various problems for ongoing and future work.

A Proof of Theorem 1

Converse part: We first show that the existence of a Slepian-Wolf code (f, g, n, M, ϵ)

[2] Taken to the extreme, when $P_Y = 0$ we have $P_Z = P_X$, hence P_Z is known, and a source code with low redundancy can be constructed.

implies the existence of a fixed rate source code for z that achieves a probability of error no larger than ϵ. The Slepian-Wolf code (f, g, n, M, ϵ) reconstructs x at the decoder with a probability of error

$$\Pr(\text{error}) = \sum_{y \in \{0,1\}^n} \Pr(y) \Pr(\text{error}|y) \leq \epsilon.$$

Therefore, there exist values of y for which $\Pr(\text{error}|y) \leq \epsilon$. Let y^* be such a value of y. We now construct a fixed rate source code for z by defining an encoder and decoder pair (h, i) based on the Slepian-Wolf pair (f, g). Let $h(z) = f(x) = f(y^* \oplus z)$ and $i(h(z)) = g(y^*, f(y^* \oplus z))$. The source encoder for z sends the index $h(z) \in \{1, \ldots, M\}$ to the source decoder, which reconstructs z by $\widehat{z} = i(h(z)) \in \{0,1\}^n$. The source code has an error, i.e., $i(h(z)) \neq z$, if and only if, given y^*, the Slepian-Wolf code (f, g, n, M, ϵ) has an error, i.e., $g(y^*, f(y^* \oplus z)) \neq y^* \oplus z$. Therefore, $\Pr(i(h(z)) \neq z) \leq \epsilon$.

We now show that a source code for z that achieves a probability of error no larger than ϵ must have a codebook size M that satisfies

$$\log(M)/n > H\left(q + \Phi^{-1}(\epsilon)\sqrt{\frac{q(1-q)}{n}}\right) + o(1/\sqrt{n}), \tag{7}$$

where we choose the $o(1)$ term in (1) such that $\Pr(z \in T^C) > \epsilon$ in (2). As discussed in Section 3.1, any set T' such that $\Pr(z \in T') \geq \Pr(z \in T)$ has cardinality $|T'|$ no smaller than $|T|$. Therefore, a source code that assigns different indices to different elements of such a set T' must have a probability of error that exceeds ϵ. But the source code that we constructed achieves a probability of error not greater than ϵ, and so we conclude that $M > |T|$.[3] By invoking Lemma 1 we have (7), where the higher order terms are all swallowed in the $o(1/\sqrt{n})$ term. Because $R_{\text{NA}}(n, \epsilon) = \log(M)/n$, (7) provides the required converse bound (3) on the non-asymptotic Slepian-Wolf rate.

Direct part: We construct a Slepian-Wolf encoder and decoder pair (f, g), where the codebook size M is chosen such that

$$M = 2^{nH\left(q+\Phi^{-1}(\epsilon)\sqrt{\frac{q(1-q)}{n}}\right)+o(\sqrt{n})}, \tag{8}$$

the significance of the $o(\sqrt{n})$ term will be clarified shortly, and the probability of error is less than ϵ. Our encoder $f(x)$ assigns x an index in the set $\{1, \ldots, M\}$ randomly, where $\Pr(f(x) = i) = 1/M$ for any $i \in \{1, \ldots, M\}$.

In order to construct our decoder $g(y, f(x))$, we must define joint typicality. We say that x and y are jointly typical if $z = x \oplus y \in T$. Our decoder determines which inputs $\widehat{x} \in \{0,1\}^n$ are jointly typical with y and are assigned the same index, i.e., $f(\widehat{x}) = f(x)$. If there is a single such sequence x', we have $g(y, f(x)) = x'$. If there are multiple or no such sequences, the decoder is incapable of reconstructing x.

We now evaluate the probability of error. The first type of error is lack of joint typicality of x and y. This relies exclusively on whether $z \in T$. We choose the $o(1)$ term in (1), which is reflected in the $o(\sqrt{n})$ term in the codebook size (8), such that $\Pr(z \in T^C) < \epsilon$ in (2). The second type of error is existence of another input $x' \neq x$ that is jointly typical with y and is assigned the same index. Using the definition of joint typicality and invoking Lemma 1, there are

$$2^{nH(t/n)+o(\sqrt{n})-0.5\log(n)+O(1)}$$

[3] Our source code may map multiple inputs to the same index, so $T' = \{z : \ i(h(z)) = z\}$ may satisfy $|T'| < M$.

input sequences that are jointly typical with y. Because $f(x)$ is assigned randomly, by choosing the $o(\sqrt{n})$ term in (8) to be "large enough", we ensure that the probability that the second type of error occurs is negligible, regardless of n. Combining the probabilities of both types of error, the probability of error of our Slepian-Wolf code is less than ϵ. $\square$

Acknowledgments

We thank Tsachy Weissman for several useful discussions. We also thank Robert Gallager, Michael Orchard, Shriram Sarvotham, Shlomo Shamai, and Prashant Singh for their comments.

References

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

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

[3] J. Wolfowitz, *Coding theorems of information theory*, 3d ed. Springer-Verlag, Berlin; New York, 1978.

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

[5] 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, March 2003.

[6] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Information Theory*, vol. IT-21, pp. 226-228, March 1975.

[7] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Information Theory*, vol. IT-47, no. 2, pp. 619-637, February 2001.

[8] R. G. Gallager, *Information theory and reliable communication*. John Wiley and Sons, New York, 1968.

[9] Y. Oohama, "Universal coding for correlated sources with linked encoders,", *IEEE Trans. Information Theory*, vol. IT-42, no. 3, pp. 837-847, May 1996.

[10] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "Refined bounds on non-asymptotic channel capacity," *in preparation*, 2004.

[11] A. Kimura and T. Uyematsu, "Weak variable-lenth Slepian-Wolf coding with linked encoders for mixed source," *IEEE Trans. Information Theory*, vol. 50, no.1, pp.183-193, January 2004.