“Real” Slepian-Wolf Codes

S. Shenvi, B. K. Dey, S. Jaggi, M. Langberg

Abstract— **We provide a novel achievability proof of the Slepian-Wolf theorem for i.i.d. sources over finite alphabets. We demonstrate that random codes that are linear over the real field achieve the classical Slepian-Wolf rate region. For finite alphabets we show that decoding is equivalent to solving an integer program. The techniques used may be of independent interest for code design for a wide class of information theory problems, and for the field of compressed sensing.**

I. Introduction

A well-known result by Slepian and Wolf in [1] characterizes the rate region for lossless source coding of distributed sources. The result demonstrated that if two (or more) sources possess correlated data, even independent encoding of the sources' data can still achieve essentially the same performance as when the sources encode jointly. This result has important implications for information theory problems as diverse as sensor networks [2], secrecy [3], and low-complexity video encoding [4]. Unfortunately for the distributed source coding problem, codes that are provably both rate optimal and computationally efficient to implement are hard to come by. Section II gives a partial history of results for the Slepian-Wolf (SW) problem.

In this work we provide novel codes that asymptotically achieve the SW rate region with vanishing probability of error. Our encoding procedure comprises of random linear operations over the real field $\mathbb{R}$, and are hence called *Real Slepian-Wolf Codes* or RSWCs. In contrast most other codes in the literature operate over appropriate finite fields $\mathbb{F}_q$. We then demonstrate that RSWCs can be used in a way that enables the receiver to decode by solving a set of integer programs. Besides being interesting in their own right as a new class of codes achieving the SW rate-region, the relation between RSWCs and IPs has some intriguing implications.

In general IPs are computationally intractable to solve. However, our code design gives us a great amount of flexibility in choosing the particular IPs corresponding to our codes. That is, we show that “almost all” RSWCs result in IPs that have “good” performance for the SW problem. But there are well-studied classes of IPs that are known to be computationally tractable to solve (for e.g., IPs corresponding to *Totally Unimodular matrices* [5]). It is thus conceivable that suitably chosen RSWCs may be decodable with low computational complexity.

At a high level, since RSWCs are linear over $\mathbb{R}$, they allow code designers to use the large toolkit of techniques available from the field of convex optimization. In contrast, since most other linear codes for the SW problem are over some finite field $\mathbb{F}_q$, decoding is usually equivalent to finding a vertex of a hypercube satisfying some combinatorial properties. Such problems are usually computationally intractable.

Also, our work has direct implications for the new field of *Compressed Sensing (CS)*. In the CS setup, N sources each generate a single real number in a manner such that the length-N sequence is *k-sparse*, i.e., can be written with at most $k \ll N$ non-zero coefficients in a predefined basis. A typical result in this setup shows that if a receiver gets $O(k\log(N))$ random linear combinations over $\mathbb{R}$ of the sources' sequence, it can, with high probability, reconstruct the source sequence exactly in a computationally efficient manner by solving a linear program. The CS setup is quite similar to that of RSWCs – the source sequence contains a large amount of redundancy, and a random $\mathbb{R}$-linear mixture of the sequence suffices for exact reconstruction via optimization techniques. There are, however, two major differences. First, RSWCs operate at information-theoretically optimal rates whereas CS codes are bounded away from such performance. Second, CS codes are computationally tractable, whereas we are currently not aware of efficient decoding techniques for RSWCs. We think this tradeoff between computational efficiency and rate-optimality is interesting and worthy of further investigation.

II. Background

Shannon's seminal source coding theorem [6] demonstrated that a sequence of discrete random variables can essentially be compressed down to the entropy of the underlying probability distribution generating the sequence. Of the many extensions sparked by this work, the Slepian-Wolf theorem [1] is the one this work builds on.

A. Slepian Wolf Theorem for i.i.d. sources [1]

Problem Statement: Two sources named Xavier and Yvonne generate two sequences of discrete random variables, $\mathbf{X} \triangleq X_1, X_2, \ldots, X_n$ over the alphabet $\mathcal{X}$, and $\mathbf{Y} \triangleq Y_1, Y_2, \ldots, Y_n$ over the alphabet $\mathcal{Y}$, respectively. The sequence $(\mathbf{X}, \mathbf{Y})$ is assumed to be i.i.d. with a joint distribution $p_{X,Y}(x,y)$ that is known in advance to both Xavier and Yvonne. The corresponding marginals are denoted by $p_X(x)$ and $p_Y(y)$. Xavier and Yvonne wish to communicate $(\mathbf{X}, \mathbf{Y})$ to a receiver Zorba. To this end Xavier uses his encoder to transmit a message that is a function only of $\mathbf{X}$ and $p_{X,Y}(x,y)$ to Zorba. Similarly, Yvonne uses her encoder to transmit a message that is a function only of $\mathbf{Y}$ and $p_{X,Y}(x,y)$ to Zorba. Zorba uses his decoder to attempt to

[0] S. Shenvi and B. K. Dey are with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India, 400 076, email: {sagarshenvi,bikash}@ee.iitb.ac.in .
S. Jaggi is with the Department of Information Engineering, Chinese University of Hong Kong, Shatin, N.T., Hong Kong, email: jaggi@ie.cuhk.edu.hk
M. Langberg is with the Computer Science Division, Open University of Israel, 108 Ravutski St., Raanana 43107, Israel, email: mikel@openu.ac.il

reconstruct $(\mathbf{X}, \mathbf{Y})$. Xavier and Yvonne's encoders and Zorba's decoder comprise a SW code C. The SW code C is said to be *near-lossless* if Zorba's reconstruction of $(\mathbf{X}, \mathbf{Y})$ is correct with a probability of error over $p_{XY}(x, y)$ that is exponentially small in the *block-length* n. The *rate-pair* (R_X, R_Y) is said to be *achievable* for the SW problem if for every $\epsilon > 0$ there exists a code C that is near-lossless, and the average (over $p_{X,Y}(x, y)$) number of bits that C requires Xavier and Yvonne to transmit to Zorba are at most $n(R_X + \epsilon)$ and $n(R_Y + \epsilon)$ respectively. The set of all rate-pairs that are achievable is called the *rate-region*.

The rate region: Slepian and Wolf's characterization of the rate-region is remarkably clean.

Theorem 1: [1] The rate region for the Slepian-Wolf problem is given by the intersection of

$$R_X \geq H(X|Y),\ R_Y \geq H(Y|X),\ R_X + R_Y \geq H(X, Y). \quad (1)$$

Here $H(X|Y)$ and $H(Y|X)$ denote the *conditional entropy* and $H(X, Y)$ denotes the *joint entropy* of (X, Y) (implicitly, over the joint distribution $p_{X,Y}(x, y)$).

B. Linear SW codes over finite fields

The SW codes in [1] have computational complexity that is exponential for both encoding and decoding. An improvement was made by [7], who showed that *random linear* encoders suffice. We briefly restate that result here, restricting ourselves to the case when $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ for simplicity.

Let $\mathbf{D_X}$ and $\mathbf{D_Y}$ be respectively $n(R_X + \epsilon) \times n$ and $n(R_Y + \epsilon) \times n$ matrices over the finite field $\mathbb{F}_2$, with each entry of both matrices chosen i.i.d. as either 0 or 1 with probability 1/2. Here ϵ is an arbitrary positive constant. Abusing notation, let $\mathbf{X}$ and $\mathbf{Y}$ also denote length-n column vectors over $\mathbb{F}_2$. Xavier and Yvonne's encoders are then defined respectively via the matrix multiplications $\mathbf{D_X X}$ and $\mathbf{D_Y Y}$, and their messages to Zorba are respectively the resulting column vectors.

We now define Zorba's decoder. For an arbitrary distribution $p_{X,Y}(x, y)$ over finite alphabets, let the *strongly ϵ-jointly typical set* $A^n_{\epsilon,p_{X,Y}}$ [8] be the set of all length-n sequences $(\mathbf{X}, \mathbf{Y})$ such that the empirical distribution induced by $(\mathbf{X}, \mathbf{Y})$ differs component-wise from $p_{X,Y}(x, y)$ by at most $\epsilon/(|\mathcal{X}||\mathcal{Y}|)$. For simplicity of notation we denote $A^n_{\epsilon,p_{X,Y}}$ as A_ϵ. Zorba checks to see if there exists a unique length-n sequence $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ satisfying two conditions. First, that $\mathbf{D_X}\hat{\mathbf{X}}$ and $\mathbf{D_Y}\hat{\mathbf{Y}}$ respectively match the messages transmitted by Xavier and Yvonne. Second, whether $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ lie within A_ϵ. If both conditions are satisfied for exactly one sequence $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$, Zorba outputs $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$, else he declares a decoding error.

Then [7] shows the following result.

Theorem 2: [7] For each rate pair (R_X, R_Y) in the region defined by (1) and sufficiently large n, with high probability over choices of $\mathbf{D_X}$ and $\mathbf{D_Y}$, the corresponding SW code is near-lossless.

Many of the SW codes in the literature build on such encoders that are linear over a finite field. Some such codes use iteratively decodable channel codes to attain performance that is empirically "good", but performance guarantees have not been proven (e.g. [9]). Other codes use recent theoretical advances in channel codes to produce near-lossless codes that achieve any point in the SW rate-region, but cannot give guarantees on computational complexity (e.g. [10]).

C. Linear codes over real fields

As mentioned in the introduction, Compressed Sensing codes operate over real (and complex) fields, and are structurally similar to the codes proposed in this work. The primary difference between the two sets of results is that our focus is on achieving information-theoretically optimal performance, whereas CS codes trade rate for lower computational complexity. Some intriguing results on CS codes can be found in [11], [12].

Concurrently, codes over reals also seem to have applications for the channel coding problem. Using significantly different techniques, Tao et al. [13] obtained channel codes that can be decoded solving a linear program (LP). Also, *lattice codes* have been shown to achieve capacity for the AWGN channel [14].

III. RSWC Model

As is usual in the SW literature [8], we focus on just the point $(H(X), H(Y|X))$ in the SW rate region. Time-sharing between this and the symmetric point $(H(X|Y), H(Y))$ enables us to achieve all points in the rate region. Thus Xavier encodes his data $\mathbf{X}$ using a classical lossless source code and Zorba decodes it losslessly. We thus discuss only Yvonne's RSWC encoder for $\mathbf{Y}$ and Zorba's corresponding decoder. In Section VI we show how to generalize our proof techniques to get codes that achieve any points in the SW rate-region without time-sharing. We consider only $\mathcal{X}$ and $\mathcal{Y}$ that are ordered *finite* subsets of $\mathbb{R}$.

RSWC Encoder: We define an $\mathbb{R}^{m\times n}$ *encoding matrix* $\mathbf{D}$. Here m is a code-design parameter to be specified later, and $\mathbf{D}$ is chosen as follows. Each component $\mathbf{D}_{ij}$ of $\mathbf{D}$ is chosen randomly from a finite set $\mathcal{D}$. More precisely, each element of $\mathbf{D}$ is chosen i.i.d. from $\mathcal{D}$ according to a distribution p_D. The set $\mathcal{D}$ can be any arbitrary finite subset of $\mathbb{R}$, and the distribution p_D can be chosen arbitrarily on $\mathcal{D}$, as long as the probability of at least two elements of $\mathcal{D}$ is non-zero. For ease of proof, we assume that p_D is zero-mean – the more general case requires only small changes in the proof details. The particular values of $\mathcal{D}$ and p_D can be chosen according to the application.

For a fixed block-length n, Yvonne's data is arranged as a column vector $\mathbf{Y} \triangleq (Y_1, Y_2, \cdots, Y_n)^T$. To encode, $\mathbf{Y}$ is multiplied by $\mathbf{D}$ to get a length-m real vector $\mathbf{U} \triangleq \mathbf{DY}$. We denote the real interval $(-n^{0.5+\epsilon}, n^{0.5+\epsilon})$ by I_q. Each component U_i of $\mathbf{U}$ is uniformly quantized by dividing I_q into steps of size $\Delta_n = 2n^{-\epsilon}$. Thus $\lceil(0.5 + 2\epsilon)\log n\rceil$ bits suffice for this quantization. Here and throughout the paper log(.) denotes the binary logarithm, and ϵ is a code-design parameter that can be used to trade off between the probability of error and the rate of the RSWC. It can be chosen as any arbitrarily small positive real number. The quantized value of U_i is denoted by $\hat{U}_i$ and the corresponding length-m quantized vector is denoted by $\hat{\mathbf{U}}$. We take $m = \lceil (n(H(Y|X) + 3\epsilon))/(0.5 \log n) \rceil$. Thus the total number of bits Yvonne transmits to Zorba equals $m\lceil (0.5 + 2\epsilon) \log n \rceil$, which for all sufficiently large n is at most $nH(Y|X) + \rho\epsilon n$ for a universal constant ρ.

RSWC Decoder: Zorba first decodes $\mathbf{X} = \mathbf{x}$. Suppose he received $\hat{\mathbf{U}} = \hat{\mathbf{u}}$ from Y. He finds a vector $\mathbf{y}$ which is strongly ϵ-jointly typical with $\mathbf{x}$, for which $\mathbf{D}\mathbf{y} \in I_q^m$, and for which $\widehat{\mathbf{D}\mathbf{y}} = \hat{\mathbf{u}}$. If there is no such $\mathbf{y}$ or there are more than one such $\mathbf{y}$ then he declares a decoding error.

The ensemble of RSWC encoder-decoder pairs described above is denoted by $C(\epsilon, n, p_{X,Y}, p_D)$. The *probability of error of* $C(\epsilon, n, p_{X,Y}, p_D)$ is defined as the probability over $p_{X,Y}$ and p_D that Zorba makes or declares a decoding error. The *rate of* $C(\epsilon, n, p_{X,Y}, p_D)$ is defined as the number of bits that Yvonne transmits to Zorba.

We are now in a position to state and prove our main results.

Theorem 3: For all sufficiently large n there are universal positive constants c, ρ, such that the probability of error and rate of $C(\epsilon, n, p_{X,Y}, p_D)$ are at most $2^{-cn^{2\epsilon}}$ and $H(Y|X) + \rho\epsilon$ respectively.

Theorem 4: For any finite alphabet $\mathcal{Y}$, the real SW encoding can done using $|\mathcal{Y}| - 1$ RSWC encoders so that the decoder can be implemented by solving $|\mathcal{Y}| - 1$ IPs.

In the rest of the paper, many different constants, independent of n, will be denoted by the same symbol 'c' for simplicity.

IV. Proof of Theorem 3

Clearly, the probability of decoding error is given by

$$P_e^n \leq P_1 + P_2 + P_3 \tag{2}$$

where P_1 is the probability that $\mathbf{D}\mathbf{Y} \notin I_q^m$, P_2 is the probability that $\mathbf{Y}$ is not strongly ϵ-jointly typical with $\mathbf{X}$, and P_3 is the probability that $\mathbf{D}\mathbf{Y} \in I_q^m$, $(\mathbf{X}, \mathbf{Y}) \in A_\epsilon$, but there is another $\mathbf{y}'$ so that $\mathbf{D}\mathbf{y}' \in I_q^m$, $(\mathbf{X}, \mathbf{y}') \in A_\epsilon$, and $\widehat{\mathbf{D}\mathbf{Y}} = \widehat{\mathbf{D}\mathbf{y}'}$.

The structure of the proof is as follows. Lemma 5 provides a concentration result on the value of each $\mathbf{D}_i\mathbf{y}$, which is used to get a bound on P_1. For $\mathbf{y} \neq \mathbf{y}'$, Lemma 6 provides an upper bound on the probability that $\widehat{\mathbf{D}_i\mathbf{y}} = \widehat{\mathbf{D}_i\mathbf{y}'}$. Lemma 7 uses Lemmas 6 to get an upper bound on P_3.

Let us define $\mathcal{D}\mathcal{Y} \triangleq \{dy | d \in \mathcal{D}, y \in \mathcal{Y}\}$. Then the following lemma gives an upper bound on the probability $Pr\{|U_i| > n^{0.5+\epsilon}\}$.

Lemma 5: If $\mathbf{D}_i$ and $\mathbf{Y}$ are as defined above, then

$$P_1' \triangleq Pr\left\{|\mathbf{D}_i\mathbf{Y}| > n^{0.5+\epsilon}\right\} \leq 2(n+1)^{|\mathcal{D}\mathcal{Y}|} 2^{-n^{2\epsilon}/2a^2 \ln 2}.$$

Proof: Let us define $P_{11} \triangleq Pr\left\{\mathbf{D}_i\mathbf{Y} > n^{0.5+\epsilon}\right\}$ and $P_{12} \triangleq Pr\left\{\mathbf{D}_i\mathbf{Y} < -n^{0.5+\epsilon}\right\}$. Then clearly $P_1' = P_{11} + P_{12}$. Let us also define $E_i \triangleq \{(d_{i1}y_1, d_{i2}y_2, \cdots, d_{in}y_n) | \sum_{j=1}^n d_{ij}y_j > n^{0.5+\epsilon}\}$. The elements $d_{ij}y_j$ take values from $\mathcal{D}\mathcal{Y}$. Let p_{yd} denote the probability mass distribution of $D_{ij}Y_j$. Then,

$$P_{11} = Pr\{E_i\} = Pr\left\{p_n | \mu_{p_n} > n^{\epsilon - 0.5}\right\}$$

Here p_n denotes the type of $(d_{i1}y_1, d_{i2}y_2, \cdots, d_{in}y_n)$ and μ_{p_n} denotes the mean of p_n. By Sanov's Theorem [8],

$$P_{11} \quad = \quad p_{yd}^n(E_i) \leq (n+1)^{|\mathcal{D}\mathcal{Y}|} 2^{-nD(p_n^* \| p_{yd})}$$

where $p_n^* = \arg\min_{p_n : \mu_{p_n} > n^{\epsilon-0.5}} D(p_n \| p_{yd})$. Now, if $a = \max\{|yd| : d \in \mathcal{D}, y \in \mathcal{Y}\}$, then $\mu_{p_n^*} > n^{\epsilon - 0.5}$ implies $|p_n^* - p|_1 > n^{\epsilon-0.5}/a \Rightarrow |p_n^* - p_{yd}|_1^2 > n^{2\epsilon-1}/a^2$. So, $D(p_n^* \| p_{yd}) \geq \frac{1}{2 \ln 2} |p_n^* - p|_1^2 > \frac{n^{2\epsilon-1}}{2a^2 \ln 2}$ (Lemma 12.6.1, [8]). So,

$$\begin{aligned} P_{11} &\leq (n+1)^{|\mathcal{D}\mathcal{Y}|} 2^{-n.n^{2\epsilon-1}/2a^2 \ln 2} \\ &= (n+1)^{|\mathcal{D}\mathcal{Y}|} 2^{-n^{2\epsilon}/2a^2 \ln 2} \end{aligned}$$

Similarly, $P_{12} \leq (n+1)^{|\mathcal{D}\mathcal{Y}|} 2^{-n^{2\epsilon}/2a^2 \ln 2}$ and thus the result follows. □

Now, $P_1 \triangleq \{|\mathbf{D}_i\mathbf{Y}| > n^{0.5+\epsilon}$ for at least one $i\} \leq \lceil n(H(Y|X + 3\epsilon)/0.5 \log n \rceil P_1'$ by union bound. Using Lemma 5, for some constant c, we have

$$P_1 \quad \leq \quad 2^{-cn^{2\epsilon}}. \tag{3}$$

For P_2, note that for any non-typical sequence $(\mathbf{x}, \mathbf{y})$, its type $p_{(\mathbf{x},\mathbf{y})}$ satisfies $|p_{X,Y} - p_{(\mathbf{x},\mathbf{y})}|_1 \geq \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}$. So, using $D(p_{X,Y} \| p_{(\mathbf{x},\mathbf{y})}) \geq \frac{1}{2 \ln 2} |p_{X,Y} - p_{(\mathbf{x},\mathbf{y})}|_1^2$ (Lemma 12.6.1, [8]) and Sanov's theorem (Theorem 12.4.1, [8]), we have

$$P_2 \quad \leq \quad (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-\frac{n\epsilon^2}{2|\mathcal{X}|^2|\mathcal{Y}|^2}\right) \leq 2^{-cn}. \tag{4}$$

for some positive constant c.

The following lemma gives, for two different $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}^n$, an upper bound on the probability that $\widehat{\mathbf{D}_i\mathbf{y}} = \widehat{\mathbf{D}_i\mathbf{y}'}$. Suppose t is the number of nonzero components in $\mathbf{y} - \mathbf{y}'$. Note that, for an i.i.d. sequence of zero mean random variables $V_1, V_2, \ldots, V_t$, the Berry-Esseen theorem [15] gives a uniform upper bound on the deviation of $W_t \triangleq \sum_{i=1}^t V_i / \sqrt{t}\sigma'$ from a zero-mean unit-variance normal distribution $\mathcal{N}(0, 1)$ as

$$|Pr\{W_t < w\} - \Phi(w)| \leq \frac{a}{\sqrt{t}}, \tag{5}$$

where σ' is the standard deviation of V_1, $\Phi(w)$ is the cumulative distribution function of $\mathcal{N}(0, 1)$, and a is a constant which depends on the distribution of V_1.

Let $p_\pm$ denote the minimum of $Pr\{D_{ij} > 0\}$ and $Pr\{D_{ij} < 0\}$. Since D_{ij} has zero mean, $p_\pm \neq 0$. Let b_y be the *smallest difference in* $\mathcal{Y}$, i.e., $b_y \triangleq \min_{y_1, y_2 \in \mathcal{Y}, y_1 \neq y_2} |y_1 - y_2|$.

Lemma 6: If $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}^n$ differ in t components,

$$Pr\{|\mathbf{D}_i(\mathbf{y} - \mathbf{y}')| < \Delta_n\} \leq \min\left(1 - p_\pm, \frac{c}{\sqrt{t}}\right)$$

for some fixed constant $c \in \mathbb{R}$.

Proof: Let us denote $(y_j - y_j')$ by α_j. Then there are t nonzero α_j, and w.l.o.g., assume that $\alpha_1, \alpha_2, \ldots, \alpha_t \neq 0$. Let us define $l \triangleq |\{y - y' | y, y' \in \mathcal{Y}, y \neq y'\}|$. Then there are at least $\tau \triangleq t/l$ elements among $\alpha_1, \alpha_2, \ldots, \alpha_t$ which are the same. Let us assume, w.l.o.g., that $\alpha_1 = \alpha_2 = \cdots = \alpha_\tau$. Then the Berry-Esseen theorem can be used on $\alpha_1 D_{i1}, \alpha_2 D_{i2}, \ldots, \alpha_\tau D_{i\tau}$ as follows.

$$
\begin{aligned}
&Pr\{|\mathbf{D}_i(\mathbf{y}-\mathbf{y}')| < \Delta_n\}\\
&= Pr\{-\Delta_n < \mathbf{D}_i(\mathbf{y}-\mathbf{y}') < \Delta_n\}\\
&= Pr\left\{-\frac{\Delta_n}{|\alpha_1|\sigma\sqrt{\tau}} < \frac{\mathbf{D}_i(\mathbf{y}-\mathbf{y}')}{|\alpha_1|\sigma\sqrt{\tau}} < \frac{\Delta_n}{|\alpha_1|\sigma\sqrt{\tau}}\right\}\\
&\le Pr\left\{-\frac{\Delta_n}{\sigma b_y\sqrt{\tau}} < \frac{\mathbf{D}_i(\mathbf{y}-\mathbf{y}')}{|\alpha_1|\sigma\sqrt{\tau}} < \frac{\Delta_n}{\sigma b_y\sqrt{\tau}}\right\}\\
&= Pr\left\{-\frac{\sum_{j=\tau+1}^{n} D_{ij}(y_j-y'_j)}{|\alpha_1|\sigma\sqrt{\tau}} - \frac{\Delta_n}{\sigma b_y\sqrt{\tau}}\right.\\
&\qquad < \frac{\sum_{j=1}^{\tau} D_{ij}(y_j-y'_j)}{|\alpha_1|\sigma\sqrt{\tau}}\\
&\qquad \left. < -\frac{\sum_{j=\tau+1}^{n} D_{ij}(y_j-y'_j)}{|\alpha_1|\sigma\sqrt{\tau}} + \frac{\Delta_n}{\sigma b_y\sqrt{\tau}}\right\} \qquad (6)\\
&\le \frac{2\Delta_n}{\sigma b_y\sqrt{\tau}\sqrt{2\pi}} + \frac{2a}{\sqrt{\tau}} \qquad = \frac{c}{\sqrt{t}} \qquad (7)
\end{aligned}
$$

Here σ^2 is the variance of d_{ij}. Equation (7) follows by using the Berry-Esseen bound (Eq. (5)) on the normalized sum in the centre of the inequality in (6). The first term in Eq. (7) is an upper bound on the probability of $\mathcal{N}(0,1)$ lying in the same range, and the second term is an upper bound on the deviation obtained by using the Berry-Esseen bound two times.

Now for $t > 0$, there is at least one j such that $y_j \neq y'_j$. Let us assume, w.l.o.g., that $y_1 \neq y'_1$. For large enough n, $\Delta_n < b_y \times \min_{d\in\mathcal{D},d\neq 0}|d|$. So, $Pr\{|\mathbf{D}_i(\mathbf{y}-\mathbf{y}')| < \Delta_n\} \le 1-p_\pm$. This can be easily checked by considering the change in the value from $\sum_{j=2}^{n} D_{ij}(y_j-y'_j)$ to $\mathbf{D}_i(\mathbf{y}-\mathbf{y}')$. □

The following lemma gives an upper bound on P_3.

Lemma 7: For constant c and large enough n, $P_3 \le 2^{-cn/\log n}$.

Proof:

$$
\begin{aligned}
P_3 &= \sum_{\substack{(\mathbf{x},\mathbf{y})\in A_\epsilon\\ \mathbf{D}\mathbf{y}\in I_q^m}} p_{X,Y}(\mathbf{x},\mathbf{y}) Pr\left\{\exists \mathbf{y}' \neq \mathbf{y} \text{ s. t. } \mathbf{D}\mathbf{y}' \in I_q^m,\right.\\
&\qquad\qquad \left.\widehat{\mathbf{D}\mathbf{y}'} = \widehat{\mathbf{D}\mathbf{y}}, (\mathbf{x},\mathbf{y}')\in A_\epsilon\right\}\\
&\le \sum_{\substack{(\mathbf{x},\mathbf{y})\in A_\epsilon\\ \mathbf{D}\mathbf{y}\in I_q^m}} p_{X,Y}(\mathbf{x},\mathbf{y}) \sum_{\substack{\mathbf{y}'\neq\mathbf{y}\\ (\mathbf{x},\mathbf{y}')\in A_\epsilon}} Pr\left\{\mathbf{D}\mathbf{y}' \in I_q^m, \widehat{\mathbf{D}\mathbf{y}'} = \widehat{\mathbf{D}\mathbf{y}}\right\}\\
&\le \sum_{(\mathbf{x},\mathbf{y})\in A_\epsilon} p_{X,Y}(\mathbf{x},\mathbf{y}) \sum_{t>0} \sum_{\substack{(\mathbf{x},\mathbf{y}')\in A_\epsilon\\ d_H(\mathbf{y},\mathbf{y}')=t}} \left(Pr\{|\mathbf{D}_1(\mathbf{y}'-\mathbf{y})| < \Delta_n\}\right)^m\\
&\le \sum_{(\mathbf{x},\mathbf{y})\in A_\epsilon} p_{X,Y}(\mathbf{x},\mathbf{y}) \sum_{t>0} \sum_{\substack{(\mathbf{x},\mathbf{y}')\in A_\epsilon\\ d_H(\mathbf{y},\mathbf{y}')=t}} \left(\min\left((1-p_\pm), \frac{c}{\sqrt{t}}\right)\right)^m\\
&\le \sum_{(\mathbf{x},\mathbf{y})\in A_\epsilon} p_{X,Y}(\mathbf{x},\mathbf{y}) \sum_{t>0} N_{\mathbf{x},\mathbf{y}}(t)\left(\min\left((1-p_\pm), \frac{c}{\sqrt{t}}\right)\right)^m
\end{aligned}
$$

where $N_{\mathbf{x},\mathbf{y}}(t) \triangleq |\{\mathbf{y}' \in \mathcal{Y}^n | (\mathbf{x},\mathbf{y}') \in A_\epsilon, d_H(\mathbf{y},\mathbf{y}') = t\}|$. Now, let us define $N(t) \triangleq \max_{\mathbf{x},\mathbf{y}} N_{\mathbf{x},\mathbf{y}}(t)$ for $t > 0$, and $t_n \triangleq \arg\max_{t>0}\left(N(t)\left(\min\left((1-p_\pm), \frac{c}{\sqrt{t}}\right)\right)^m\right)$. The subscript in t_n is to emphasize that it is a function of n. Then clearly,

$$
P_3 \quad \le \quad nN(t_n)\left(\min\left((1-p_\pm), \frac{c}{\sqrt{t_n}}\right)\right)^m .
$$

Now for any $\delta \le \frac{\epsilon}{2(H(Y|X)+3\epsilon)}$, we consider two regimes: (1) $t_n > n^{1-\delta}$ and (2) $t_n \le n^{1-\delta}$. In the first regime, we use the bound $N(t_n) \le 2^{n(H(Y|X)+2\epsilon)}$ and $Pr\{|\mathbf{D}_i(\mathbf{y}-\mathbf{y}')| < \Delta_n\} \le \frac{c}{\sqrt{t}}$, and get, for large enough n,

$$
\begin{aligned}
\log(P_3) \quad &\le \quad \log n + \log N(t_n)\\
&\qquad -\frac{n(H(Y|X)+3\epsilon)}{0.5\log n}((0.5-0.5\delta)\log n - \log c)\\
&= \quad n(H(Y|X)+2\epsilon) - n(H(Y|X)+3\epsilon)(1-\delta)\\
&\qquad +n\frac{(H(Y|X)+3\epsilon)}{0.5\log n}c + \log n\\
&= \quad -n\left(\epsilon - \delta(H(Y|X)+3\epsilon)\right)\\
&\qquad +n\left(\frac{(H(Y|X)+3\epsilon)}{0.5\log n}c + \frac{\log n}{n}\right).
\end{aligned}
$$

Now, using $\delta \le \frac{\epsilon}{2(H(Y|X)+3\epsilon)}$ and $\left(\frac{(H(Y|X)+3\epsilon)}{0.5\log n}c + \frac{\log n}{n}\right) < \epsilon/4$ for sufficiently large n, we get $\log(P_3) \le -n\epsilon/2 + n\epsilon/4 = -n\epsilon/4$. In the regime $t_n < n^{1-\delta}$, we use the bounds $N(t_n) < (|\mathcal{Y}|-1)^{t_n}\binom{n}{t} < (|\mathcal{Y}|n)^{t_n}$, and $Pr\{|\mathbf{D}_i(\mathbf{y}-\mathbf{y}')| < \Delta_n\} \le 1-p_\pm$ to get

$$
\begin{aligned}
\log(P_3) \quad &\le \quad \log n + t_n\log n + t_n\log|\mathcal{Y}| - \frac{cn(H(Y|X)+3\epsilon)}{\log n}\\
&\le \quad \log n + n^{1-\delta}\log n + n^{1-\delta}\log|\mathcal{Y}| - \frac{cn}{\log n},
\end{aligned}
$$

where c in the above two lines are two different constants.. Now, for large enough n, $(\log n)^2 < \frac{c}{3}n^\delta \Rightarrow \log n < \frac{cn^\delta}{3\log n}$. Also, for large enough n, $n^{-\delta}\log|\mathcal{Y}| < \frac{c}{3\log n}$ for some constant c. So, $\log(P_3) \le \log n - \frac{cn}{3\log n} \le -\frac{cn}{\log n}$ for large enough n and for some constant c.

Since $\frac{cn}{\log n} < n\epsilon/4$ for large enough n, the result follows. □

From Equations (2), (3), (4), and Lemma 7, we have, for large enough n, $P_e^n \le 3.P_1 \le 2^{-cn^{2\epsilon}}$, for a constant c, thus completing the proof of Theorem 3. □

V. Proof of Theorem 4

We first show that for $\mathcal{Y} = \{0,1\}$ the decoding of our scheme can be done via the solution of an IP (rather than by typicality decoding as in Section IV). For typical $\mathbf{x} = x_1,\ldots,x_n$ decoded by Zorba and $x \in \mathcal{X}$, let $I_x = \{i | x_i = x\}$. The constraint $(\mathbf{x},\mathbf{y}) \in A_\epsilon$ can be phrased as the linear constraints

$$
p(1,x) - \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \le \frac{1}{|I_x|}\sum_{i\in I_x} y_i \le p(1,x) + \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}, \quad \forall x \in \mathcal{X}
$$

Also, the constraints $\widehat{\mathbf{D}\mathbf{y}} = \hat{\mathbf{u}} = \hat{u}_1,\ldots,\hat{u}_n$ can be written as

$$
\hat{u}_i - \Delta/2 \le \mathbf{D}_i\mathbf{y} \le \hat{u}_i + \Delta/2, \quad \forall i = 1,\ldots m
$$

Finally we add the 'integrality' constraints, i.e., $\mathbf{y} \in \mathcal{Y}^n$.

For arbitrary finite alphabets $\mathcal{Y}$, Yvonne and Zorba perform $|\mathcal{Y}|-1$ encoding and decoding stages, each of which involves IP decoding of a binary vector. A sketch follows.

Let $y(1), \ldots, y(|\mathcal{Y}|)$ denote the distinct values of $\mathcal{Y}$. In the first stage, instead of encoding $\mathbf{y}$ directly, Yvonne uses $C(\epsilon, n, p^1_{X,Y}, p_D)$ to encode the vector $f^1(\mathbf{y})$. Here the vector $f^1(\mathbf{y})$ equals 1 in the locations that $\mathbf{y}$ equals $y(1)$ and equals 0 otherwise, and $p^1_{X,Y}$ is the corresponding induced distribution $p_{X,f^1(Y)}$ defined on $\mathcal{X} \times 2$. Since $f^1(\mathbf{y})$ is a binary vector, Zorba can use the IP decoding described above, and therefore can retrieve the locations where $\mathbf{y}$ equals $y(1)$. Inductively, in the ith stage, Yvonne uses $C(\epsilon, n(i), p^i_{X,Y}, p_D)$ to encode the vector $f^i(\mathbf{Y})$. Here $n(i)$ equals the number of locations whose values are still undetermined before the i^{th} stage, i.e., $n(i)$ equals $|\{j|\mathbf{y}_j \geq y(i)\}|$. The length-$n(i)$ vector $f^i(\mathbf{y})$ is obtained by first throwing away the locations in $f^{i-1}(\mathbf{y})$ that equalled 1, and then marking the remaining locations 1 iff the corresponding locations in $\mathbf{y}$ equals $y(i)$. The distribution $p^i_{X,Y}$ is the corresponding induced distribution $p_{X,f^i(\mathbf{y})}$ defined on $\mathcal{X} \times 2$. At each stage, Zorba can use the IP decoding described above, and therefore can retrieve the locations where $\mathbf{y}$ equals $y(i)$.

The overall probability of error is at most the sum of the probabilities of error of each stage, and is therefore still exponentially small in n. Since there is a bijection between $\mathbf{Y}$ and the set of random vectors $\{f^i(\mathbf{Y})\}_{i=1}^{|\mathcal{Y}|-1}$, the conditional entropy of $\mathbf{Y}$ equals the conditional entropy of $\{f^i(\mathbf{Y})\}_{i=1}^{|\mathcal{Y}|-1}$. Repeatedly applying Theorem 3 implies that the overall rate of this multistage RSWC differs from $H(Y|X)$ by at most $C'\epsilon$, where C' is a universal constant. □

VI. Real SW coding without timesharing

Any rate pair in the SW rate-region (1) can also be directly achieved by real codes without timesharing between the schemes achieving the rate pairs $(H(X|Y), H(Y))$ and $(H(X), H(Y|X))$. Let (R_1, R_2) be a rate pair in the SW rate region. Let $m_1 = \lceil (n(R_1 + 3\epsilon))/(0.5 \log n) \rceil$ and $m_2 = \lceil (n(R_2 + 3\epsilon))/(0.5 \log n) \rceil$. Xavier choses a random $m_1 \times n$ encoder matrix $\mathbf{D}_1$ over $\mathcal{D}$ and Yvonne choses a random $m_2 \times n$ encoder matrix $\mathbf{D}_2$ over $\mathcal{D}$. Though the matrices could be chosen over different finite subsets of $\mathbb{R}$, we take the same subset for simplicity. Xavier's encoder encodes the n length vector $\mathbf{X}$ by quantizing $\mathbf{U}_1 \triangleq \mathbf{D}_1\mathbf{X}$ in the range $I_q^{m_1}$ with per dimension step size $\Delta_n = 2n^{-\epsilon}$. Similarly Yvonne's encoder encodes the n length vector $\mathbf{Y}$ by quantizing $\mathbf{U}_2 \triangleq \mathbf{D}_2\mathbf{Y}$ in the range $I_q^{m_2}$ with per dimension step size $\Delta_n = 2n^{-\epsilon}$. Suppose $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{U}}_2$ are respectively the quantized vectors. The decoder finds a unique strongly ϵ-jointly typical pair $(\mathbf{x}, \mathbf{y})$ so that $\widehat{\mathbf{D}_1\mathbf{x}} = \widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{D}_2\mathbf{y}} = \widehat{\mathbf{U}}_2$. If there are no or multiple such pairs the decoder declares error. Analogously to Theorem 3, the probability of error and the total rate $R_1 + R_2$ can be bounded from above by $2^{-cn^{2\epsilon}}$ and $H(X,Y) + \rho\epsilon$ respectively. Details omitted here can be found in [16].

VII. Conclusion

The RSWCs analyzed here provide a novel achievability proof of the SW theorem. Perhaps just as importantly, they demonstrate the intriguing possibility of design of information-theoretic codes via convex optimization techniques. For instance, since decoding RSWCs is equivalent to solving an optimization problem, it is natural to consider similar "real" codes for problems where some function of the code simultaneously needs to be optimized. We are currently investigating the performance of RSWCs under more structured choices of encoding matrices, with the hope of obtaining codes for which IP decoding is equivalent to LP decoding, and is therefore computationally tractable.

VIII. Acknowledgments

The authors gratefully acknowledge support from the CUHK direct grant, the CU-MS-JL grant, and a grant from the Bharti Center for Communication. We would also like to thank D. Manjunath for fruitful discussions, and the reviewers for valuable comments.

References

[1] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory,* 19:471–480, July 1973.

[2] J. Kusuma S. Pradhan and K. Ramchandran. Distributed compression in a dense microsensor network. *IEEE Signal Processing Magazine,* 19:51–60, March 2002.

[3] I. Csiszár and P. Narayan. Common randomness and secret key generation with a helper. *IEEE Transactions on Information Theory*, 46(2):344–366, March 2000.

[4] R. Puri and K. Ramchandran. Prism: a new robust video coding architecture based on distributed compression principles. In *Proceedings of the Allerton Conference on Communications, Control, and Computing*, October 2002.

[5] A. J. Hoffmann. The role of unimodularity in applying linear inequalities to combinatorial theorems. *Annals of Discrete Mathematics*, 4:73–84, 1979.

[6] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[7] I. Csiszar. Linear codes for sources and source networks: error exponents, universal coding. *IEEE Transactions on Information Theory*, 28(4):585–592, 1982.

[8] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[9] J. Garcia-Frias and Y. Zhao. Compression of correlated binary sources using turbo codes. *IEEE Communication Letters*, pages 417–419, October 2001.

[10] M. Médard M. Effros T. P. Coleman, A. H. Lee. On some new approaches to practical slepian-wolf compression inspired by channel coding. In *Proceedings of the Conference on Data Compression*, page 282, March 2004.

[11] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[12] J. Romberg E. Candès and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.

[13] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[14] R. Urbanke and B. Rimoldi. Lattice codes can achieve capacity on the awgn channel. *IEEE Transactions on Information Theory*, 44(1):273–278, 1998.

[15] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume II (2nd ed.).* John Wiley & Sons, New York, 1972.

[16] S. Shenvi, B. K. Dey, S. Jaggi, and M. Langberg. "real" slepian-wolf codes. Technical report, Available at http://www.ee.iitb.ac.in/wiki/faculty/bikash.