Symmetric and a-symmetric Slepian-Wolf codes with systematic and non-systematic linear codes

Nicolas Gehrig, *Student Member, IEEE,* and Pier Luigi Dragotti, *Member, IEEE*

***Abstract*— We propose a constructive approach for distributed source coding of correlated binary sources using linear channel codes that can achieve any point of the Slepian-Wolf achievable rate region. Our approach is very intuitive and can be used with systematic and non-systematic linear codes. Moreover, the proposed coding strategy can easily be extended to the case of more than two sources.**

***Index Terms*— Distributed source coding, Slepian-Wolf theorem, symmetric encoding, linear channel codes.**

I. Introduction

THE Slepian-Wolf theorem [1] states that separate encoding of the outputs of two correlated sources can be as efficient as joint encoding, assuming that the two compressed signals can be jointly decoded. The achievable rate region for such a system is given by: $R_X \geq H(X|Y)$, $R_Y \geq H(Y|X)$ and $R_X + R_Y \geq H(X,Y)$.

Although this theoretical result has been known for about three decades, it is only recently that practical coding approaches have been proposed. In [2], a first coding technique using channel coding principles was introduced. Practical designs mainly based on Turbo and LDPC codes have since been presented in several other papers (see [3], [4], [5] for example). Most of these approaches focus on the asymmetric scenario, where one of the two sources is transmitted perfectly to the receiver.

For practical applications, it might be necessary to have more flexibility in the repartition of the bit-rates between the encoders. In [6], Pradhan and Ramchandran proposed a technique based on their original work (DISCUS [2]) in order to achieve any point of the Slepian-Wolf achievable rate region. Their method creates two sub-codes of a single channel code by splitting the original generator matrix in two. Each encoder uses then one of these sub-codes to encode its data.

In this letter, we propose a constructive approach that allows for a flexible repartition of the transmission rates between the encoders. Our technique uses a single linear channel code that can be non-systematic. The performance of our approach depends only on the quality of the channel code used. Actually, the two correlated sources can be seen respectively as the input and the output of a certain channel used to model their correlation. We refer to this virtual channel as the *correlation channel* of the two sources. If we can find a code that achieve the capacity of this *correlation channel*, then our distributed source coding approach can reach the Slepian-Wolf bound and is therefore optimal.

Notice that similar approaches have recently been proposed in [7], [8] and [9]. Although our own approach is relatively similar in spirit to the one in [7], our scheme can also be used with non-systematic codes, whereas their technique can only be used with systematic codes. Note that good capacity achieving LDPC codes are usually non-systematic. In [8] and [9], iterative decoding procedures are proposed in order to decode the two correlated blocks simultaneously. Their main strategy is to apply the standard sum-product algorithm (message-passing decoding [10]) on an extended factor graph, corresponding to two standard LDPC decoders connected through correlation nodes modeling the joint distribution between the sources. In Section III, we show that our approach does not require the use of such extended graphs since our methods only needs to decode one single block (the difference pattern). A standard iterative decoding scheme similar to the one proposed in [5] can therefore be used in our case. Our approach can thus be seen as an intuitive extension of the asymmetric approach proposed in [5] in order to achieve the entire Slepian-Wolf achievable rate region.

II. A simple example with the Hamming $(7,4)$ code

We consider the Hamming $(7,4)$ code $\mathcal{C}$ whose parity check matrix is given by:

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}. \quad (1)$$

We know that a codeword x belongs to the Hamming code $\mathcal{C}$ if and only if its syndrome is equal to zero:

$$s_x = \mathbf{H}x^T = 0 \iff x \in \mathcal{C}. \quad (2)$$

The minimum distance between any two of the 16 codewords of the Hamming code is three. This code is therefore able to correct up to one bit error per codeword. Assume e_i is the error pattern corresponding to an error at bit position i (e_i has six 0s and one 1 at position i). We define $y = x \oplus e_i$ where $\oplus$ corresponds to the binary addition. This codeword y does clearly not belong to $\mathcal{C}$ since its distance from x is equal to one. Its syndrome is given by:

$$s_y = \mathbf{H}y^T = \mathbf{H}(x \oplus e_i)^T = \mathbf{H}x^T \oplus \mathbf{H}e_i^T = \mathbf{H}e_i^T. \quad (3)$$

We can therefore see that the syndrome of an erroneous codeword does not depend on the original codeword but only on the error pattern. This means that if we change the i^{th} bit

The authors are with the Communications and Signal Processing Group, Electrical and Electronic Engineering Department, Imperial College London, Exhibition Road, London SW7 2AZ, UK (e-mails: {nicolas.gehrig, p.dragotti}@imperial.ac.uk).

of all the codewords of $\mathcal{C}$, this produces 16 new codewords, all having syndrome $\mathbf{H}e_i^T$. This new set of codewords is called *coset number* i and has the same properties as C (coset 0), that is, the minimum distance between any two codewords is still three. All the 2^7 possible 7-bit blocks are thus distributed in 8 distinct cosets. Notice that this Hamming $(7,4)$ code has a particular structure such that the syndrome of an erroneous codeword gives the binary representation of the error position, or similarly, the coset number.

Consider now two discrete memoryless uniformly distributed 7-bit binary random variables x and y, correlated such that their Hamming distance is at most one ($d_H(x,y) \leq 1$). Assume that x and y belong to cosets i and j respectively. The difference between x and y is given by the error pattern $e_k = x \oplus y$ (x and y differs at position k). We know that the syndromes of x and y are given by $s_x = \mathbf{H}x^T = \mathbf{H}e_i^T$ and $s_y = \mathbf{H}y^T = \mathbf{H}e_j^T$ respectively. We can see that:

$$s_k = \mathbf{H}e_k^T = \mathbf{H}(x \oplus y)^T = \mathbf{H}x^T \oplus \mathbf{H}y^T = s_x \oplus s_y. \quad (4)$$

This result shows that knowing only the syndromes of x and y, we can retrieve the syndrome of their difference pattern and therefore, the bit position where they differ.

Our coding technique can now be presented as follows: Assume the following block representations for x, y and $\mathbf{H}$:

$$x = [x_a \quad x_b] \qquad y = [y_a \quad y_b] \qquad \mathbf{H} = [\mathbf{H}_a \quad \mathbf{H}_b] \quad (5)$$

where the first and the second blocks are of length 4 and 3 respectively. The syndromes of x and y are computed at their respective encoders as: $s_x = \mathbf{H}x^T = \mathbf{H}_a x_a^T \oplus \mathbf{H}_b x_b^T$ and $s_y = \mathbf{H}y^T = \mathbf{H}_a y_a^T \oplus \mathbf{H}_b y_b^T$. Encoder 1 transmits s_x together with a subset of x_a. Encoder 2 transmits s_y together with the subset of y_a which is complementary to the one chosen by the first encoder. For example, as presented in Figure 1, the encoder 1 could send: $[x_1 \quad x_2 \quad s_x^T]$ and encoder 2: $[y_3 \quad y_4 \quad s_y^T]$.

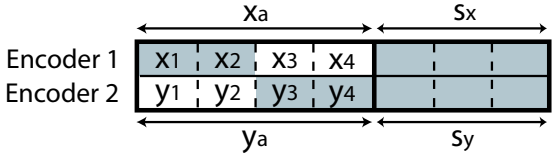

Fig. 1. Example of distributed source coding of two correlated 7-bit blocks. Only the gray squares are transmitted.

At the decoder, the syndrome of the difference pattern between x and y is obtained by computing the sum of the two syndromes $s_x \oplus s_y$. Using the corresponding error pattern, the missing bits of x_a and y_a can easily be retrieved. Finally, x_b and y_b are obtained as: $x_b^T = \mathbf{H}_b^{-1}(s_x \oplus \mathbf{H}_a x_a^T)$ and $y_b^T = \mathbf{H}_b^{-1}(s_y \oplus \mathbf{H}_a y_a^T)$.

Since x and y are uniformly distributed, we have: $H(x) = H(y) = 7$ bits. We know that y can take 8 different equiprobable values for a specific x. Hence, $H(y|x) = H(x|y) = 3$ bits. The joint entropy of x and y is therefore equal to $H(x,y) = H(x) + H(y|x) = 10$ bits. Our coding scheme uses 6 bits to send the two syndromes and a total of 4 bits to send the two complementary subsets of x_a and y_a and is therefore optimal.

III. Constructive approach using any linear channel code

Assume we have an (n,k) binary linear code $\mathcal{C}$ with parity check matrix $\mathbf{H}$ in its reduced form such that: $\mathbf{H} = [\mathbf{H}_1 \quad \mathbf{H}_2]$, where $\mathbf{H}_1$ and $\mathbf{H}_2$ are of size $(n-k \times k)$ and $(n-k \times n-k)$ respectively. Assume without loss of generality that $\mathbf{H}_2$ is non singular. Notice that if the code is systematic, we know that the generator matrix $\mathbf{G}$ is of the form: $\mathbf{G} = [\mathbf{I}_k \quad \mathbf{P}]$, where $\mathbf{P}$ is of size $(k \times n-k)$. The parity check matrix $\mathbf{H}$ has to satisfy: $\mathbf{G}\mathbf{H}^T = \mathbf{0}$ and can therefore be given by: $\mathbf{H} = [-\mathbf{P}^T \quad \mathbf{I}_{n-k}]$. $\mathbf{H}_2$ is thus simply the identity matrix in that case.

Assume $\mathcal{C}$ is able to correct up to M errors per n-bit code block. We know that the following relation must hold:

$$2^{n-k} \geq \sum_{j=0}^{M} \binom{n}{j} \qquad \text{(sphere packing bound).} \quad (6)$$

This code $\mathcal{C}$ generates 2^{n-k} cosets each containing 2^k codewords of length n. We know that x belongs to the coset i, such that $s_x = \mathbf{H}x^T = \mathbf{H}e_i^T$ (e_i is the coset leader of coset number i, i.e., the codeword with minimum weight).

Let x_i be a binary block of length n represented as: $x_i = [a_i \quad b_i \quad q_i]$, where a_i, b_i and q_i are of length k_1, k_2 and $n-k$ respectively (k_1 and k_2 are chosen such that their sum is equal to k). The syndrome of x_i is defined as: $s_i = \mathbf{H}_1[a_i \quad b_i]^T \oplus \mathbf{H}_2 q_i^T$.

Consider now two n-bit blocks x_1 and x_2, correlated such that their Hamming distance $d_H(x_1, x_2)$ is at most m (Assume that $M \geq m$). Our distributed coding strategy consists in sending only $[a_1 \quad s_1^T]$ and $[b_2 \quad s_2^T]$ from the encoders 1 and 2 respectively. The transmission bit-rates are therefore given by: $R_1 = n - k_2$ bits and $R_2 = n - k_1$ bits, corresponding to a total of $2n - k$ bits.

At the receiver, we let e_d correspond to the "difference pattern" between x_1 and x_2 as: $e_d = x_1 \oplus x_2$. We know that the syndrome of e_d is given by $s_d = \mathbf{H}e_d^T = \mathbf{H}(x_1^T \oplus x_2^T) = s_1 \oplus s_2$. We can now retrieve the error pattern e_d corresponding to this syndrome s_d using one of the following techniques: If the code is not too large, a simple lookup table storing the corresponding pattern error for each possible syndrome can be used. For larger code, an iterative method has to be used. Using an iterative decoding scheme such as the one proposed in [5], we can recover e_d as the closest codeword to the all zero sequence satisfying the syndrome s_d. Notice that this iterative decoding approach is particularly suited for LDPC codes which are amongst the best block codes known for memoryless channels [10].

Knowing the difference pattern e_d, the missing bits of the k first bits of x_1 and x_2 are easily obtained as: $[a_2 \quad b_1] = [a_1 \quad b_2] \oplus e_d^k$, where e_d^k corresponds to the k first bits of e_d.

We know that the syndrome of x_1 corresponds to: $s_1 = \mathbf{H}_1[a_1 \quad b_1]^T \oplus \mathbf{H}_2 q_1^T$. Let z_1 be defined as: $z_1 = s_1 \oplus \mathbf{H}_1[a_1 \quad b_1]^T$. We can now retrieve q_1 by computing: $q_1^T = \mathbf{H}_2^{-1} z_1$. Notice that $\mathbf{H}_2^{-1}$ can be obtained using Gaussian elimination and that, if $\mathcal{C}$ is systematic, we can choose $\mathbf{H}$ such that $\mathbf{H}_2 = \mathbf{I}$ and $q_1 = z_1^T$. The inversion of $\mathbf{H}_2$ is actually done only once (off-line) and does not introduce extra complexity to the decoding phase. This inversion is the one that is usually

done in order to compute the generator matrix from a parity check matrix.

Knowing q_1, we have now completely recovered x_1 and we can easily obtain x_2 as $x_2 = x_1 \oplus e_d$. We can summarize our coding approach as follows:

*Proposition 1: A*ssume X and Y are two binary sequences of length n, correlated such that their Hamming distance is at most m. Consider an (n, k) linear channel code $\mathcal{C}$ that can correct up to $M \geq m$ errors per n-bit block. The following distributed coding strategy uses a total of $2n-k$ bits to encode the two sequences and is sufficient to allow for a perfect reconstruction of them at the decoder:

- Send the syndromes of X and Y from their respective encoders.
- Send only complementary subsets of their first k bits.

In terms of performance, we can say that the ability of our distributed source coding technique to work close to the Slepian-Wolf bound only depends on the quality of the channel code used. More specifically, if X and Y are uniformly distributed and $p(Y|X)$ is the transition probability, then the closer the channel code $\mathcal{C}$ gets to the capacity of the binary channel $p(Y|X)$, the closer our system gets to the Slepian-Wolf bound. The design of capacity achieving channel codes, however, is beyond the scope of this paper.

Since our decoding strategy can use an iterative decoding approach similar to the one proposed in [5], similar performances in terms of decoding error probability are expected if the same linear code is used by both systems. Our work can thus be seen as an intuitive extension of the work in [5] in order to cover the full Slepian-Wolf achievable rate region without any performance loss.

We have run some simulations and we have obtained numerical evidence that our approach presents similar performances than the one proposed in [5].

IV. Generalization to more than two sources

The approach of the previous section can be extended to any number of correlated sources (see Figure 2), as indicated in the following proposition:

*Proposition 2: A*ssume $x_1, \ldots, x_L$ are L binary sequences of length n correlated such that the Hamming distance between two consecutive sequences is at most m (i.e., $d_H(x_i, x_{i+1}) \leq m$ for $i = 1, \ldots, L-1$). Consider an (n, k) linear channel code $\mathcal{C}$ that can correct up to $M \geq m$ errors per n-bit block. The following distributed coding strategy uses a total of $n + (L-1)(n-k)$ bits to encode the L sequences and is sufficient to allow for a perfect reconstruction of all of them at the decoder:

- From each encoder, send the syndrome s_i of the corresponding block x_i.
- Send only complementary subsets of their first k bits such that each bit position is sent from only one encoder.

At the decoder, the $L-1$ difference patterns can be recovered from the L syndromes, allowing then to complete the first k bits of each sequence.

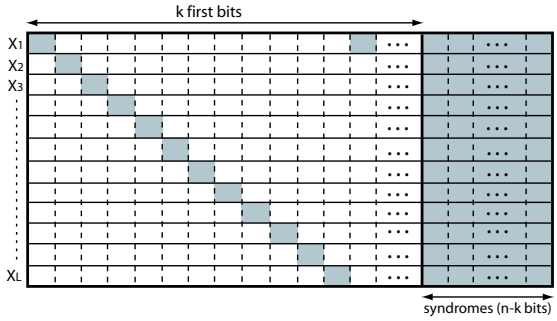

Fig. 2. Our encoding strategy for L correlated binary sources. Each encoder sends the syndrome and a subset of the first k bits of their input block.

The decoding method used here is similar to the one presented in the previous section. The difference pattern of each pair of consecutive blocks is retrieved by running the standard iterative decoding method with the sum of the two syndromes. Knowing all the difference patterns and having received complementary subsets of the first k bits, the first k bits of each block can then be recovered. Finally, each original block $x_i := [a_i \quad q_i]$ $(i = 1, \ldots, L)$ is completed by recovering its last $n-k$ bits as:

$$q_i^T = \mathbf{H}_2^{-1}(s_i \oplus \mathbf{H}_1 a_i^T). \tag{7}$$

It is possible to show that this coding strategy is optimal for some particular cases.

References

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, Jul 1973.

[2] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," in *IEEE International Conference on Data Compression (DCC)*, 1999, pp. 158–167.

[3] J. Garcia-Frias, "Compression of correlated binary sources using turbo codes," *IEEE Communications Letters*, vol. 5, no. 10, pp. 417–419, October 2001.

[4] A. Aaron and B. Girod, "Compression with side information using turbo codes," in *IEEE International Conference on Data Compression (DCC)*, April 2002, pp. 252–261.

[5] A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes," *IEEE Communications Letters*, vol. 6, no. 10, pp. 440–442, October 2002.

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

[7] V. Stankovic, A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Design of Slepian-Wolf codes by channel code partitioning," in *IEEE International Conference on Data Compression (DCC)*, March 2004.

[8] D. Schonberg, S. S. Pradhan, and K. Ramchandran, "Distributed code constructions for the entire Slepian-Wolf rate region for arbitrarily correlated sources," in *IEEE International Conference on Data Compression (DCC)*, March 2004.

[9] T. P. Coleman, A. H. Lee, M. Medard, and M. Effros, "On some new approaches to practical Slepian-Wolf compression inspired by channel coding," in *IEEE International Conference on Data Compression (DCC)*, March 2004.

[10] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 599–618, February 2001.