

Nonlinear Sparse-Graph Codes for Lossy Compression of Discrete Nonredundant Sources

Ankit Gupta

Department of Electrical Engineering
Princeton University
Princeton, NJ 08540
ankitg@princeton.edu

Sergio Verdú

Department of Electrical Engineering
Princeton University
Princeton, NJ 08540
verdu@princeton.edu

Abstract—We propose a scheme to implement lossy data compression for discrete equiprobable sources using block codes based on sparse matrices. We prove asymptotic optimality of the codes for a Hamming distortion criterion. We also present a sub-optimal decoding algorithm, which has near optimal performance for moderate blocklengths.

I. INTRODUCTION

The duality between source and channel coding has been recognized since Shannon [1]. This duality exists both in the sense of evaluating the capacity and rate distortion functions [2], as well as optimal coding schemes [3] [4]. In particular error correcting codes (used for data transmission) can be used for source coding. In [5] it is shown that linear error correcting codes may be used for lossless compression of memoryless sources with the source considered as an error pattern for the channel. This compressor is based on the $m \times n$ parity check matrix \mathbf{H} which maps \mathbf{s} the source vector to the vector $\mathbf{H}\mathbf{s}$. At the output the maximum likelihood decoder selects $\hat{\mathbf{s}}$ such that $\hat{\mathbf{s}} = g_H(\mathbf{H}\mathbf{s})$, where g_H is the maximum likelihood syndrome decoder for the error correcting code. Using this scheme we obtain a compression ratio $R = \frac{m}{n}$. However it may happen that $\mathbf{s} \neq \hat{\mathbf{s}}$ with probability $\epsilon > 0$. Thus this coding scheme is almost lossless with a block error probability of ϵ . It may be verified easily that the block error probabilities of the error correcting code and the corresponding data compressor are identical. It can also be shown through a random coding argument that linear encoding does not incur any loss of optimality for memoryless sources [4]. Using the foregoing paradigm, if we have a capacity achieving code for the binary symmetric channel then a Bernoulli source with parameter p can be compressed at a rate $h(p)$ (where $h(\cdot)$ is the binary entropy), with the probability of block error $\epsilon \rightarrow 0$ as $n \rightarrow \infty$. Practical universal lossless compressors with linear complexity based on linear sparse graph codes have been proposed in [6] and [7] and shown to have performance comparable to or better than existing data compressors such as Lempel-Ziv or Context-Tree Weighting for a wide variety of sources.

We now turn to the duality between lossy source coding and channel coding. One way to construct a lossy n -to- k

compressor is to use the decoder of an (n, k) channel code. For example, in the case of a binary source with Hamming distortion, a natural (albeit not always computationally feasible) choice of the compressor is to select the k -bit index of the codeword closest (in Hamming distance) to the input. The decompressor is simply the channel encoder. The method just outlined can be shown to achieve the rate distortion function. Consider the memoryless source S with distribution P_S , and a distortion function $d(\cdot, \cdot)$. The rate distortion function for this source is given by

$$R(d) = \min_{\substack{P_{\hat{S}|S} \\ d(\hat{S}; S) \leq d}} I(\hat{S}; S). \quad (1)$$

Solving (1) we obtain $P_{\hat{S}|S}$ (corresponding to $R(d)$) and form a channel with this transition probability. Pick a codebook for channel coding by sampling codewords randomly from the distribution $P_{\hat{S}}$ with rate $R(d) + \gamma$ (where $\gamma > 0$, is arbitrary), that achieves a block error rate equal to ϵ . This codebook can also be used to compress a n length source distributed according to P_S . To compress the source using this channel code we use the (strongly typical) channel decoder on the source realization \mathbf{s} and locate a valid codeword in the codebook. The codeword can then be represented with $n(R(d) + \gamma)$ bits. The block error rate of this source code is defined as the probability that the codeword thus located does not satisfy the distortion constraint $d(\hat{S}; S) \leq d$. Suppose with this codebook we achieve a block error probability for source coding equal to $\hat{\epsilon}$. Then it is shown in [4] that $\hat{\epsilon} = 1 - \epsilon + 2\gamma$. However since we are transmitting at a rate greater than $I(\hat{S}; P_{S|\hat{S}})$ the channel coding theorem ensures that $\epsilon \rightarrow 1$, therefore $\hat{\epsilon} \rightarrow 2\gamma$ and we achieve the rate distortion limits asymptotically.

Consider the problem of compressing the binary symmetric source with a Hamming distortion criteria. The dual of this problem is transmission over the binary symmetric channel with crossover probability d and sampling the input codewords from the binary symmetric distribution. Then an optimum decoder for such a random code with rate $R = 1 - h(d) + \gamma$ would be able to compress the binary symmetric source within average distortion d . It was also shown independently by Berger and Goblick in [3] and [8] respectively that there

This work was partially supported by the National Science Foundation under Grant CCR-0312839.

exists a linear code for compressing the binary symmetric source. It may be conjectured that LDPC codes are optimal for compressing the binary symmetric source with a Hamming distortion criterion (because of their random construction). It was shown in [9] that MacKay's random ensemble [10] is asymptotically optimal for this problem (Further, a scheme using LDPC codes was shown asymptotically optimal for compressing the discrete memoryless source with Hamming distortion in [11]). In this ensemble the number of ones in the \mathbf{H} matrix grows super-linearly with the blocklength n . Unfortunately the belief propagation decoder fails when used as a lossy encoder for these codes. Furthermore a polynomial complexity (possibly suboptimal) encoder with near optimal performance has not been found for this problem. In the absence of a practical encoder with acceptable performance, other researchers have proposed the use of the dual of LDPC viz. low density generator matrix codes (LDGM) for this problem. This approach was followed in [12] substituting parity check equations by more general Boolean operations, and in [13] using linear LDGM codes. Both [12] and [13] showed excellent empirical performance. However the asymptotic optimality of LDGM codes for this problem has not been shown. Another approach using a LDPC-LDGM hybrid code with finite check degrees is proven asymptotically optimal in [14] but a viable near optimum encoding algorithm is not known. Thus so far no codes which are asymptotically optimal under maximum likelihood encoding and have suboptimal encoding algorithms with near optimal performance have been found for lossy compression.

In this paper we propose nonlinear codes, based on LDGM matrices, which are asymptotically optimal for compressing the nonredundant discrete source with a Hamming distortion criterion, which includes the binary symmetric source as a special case. We also provide a suboptimal decoding algorithm for these codes, which has excellent empirical performance, even at moderate blocklengths. For the discrete equiprobable (q -ary source) the rate distortion function with a Hamming distortion criterion is given as

$$R(d) = \log_2(q) - d \log_2(q-1) - h(d). \quad (2)$$

Our code design is an intermediate on a continuum of block codes with the linear construction and the random codebook as the two extremes. These codes and encoding algorithms can also be extended for the more general case of compressing the discrete memoryless source with a separable distortion criterion (with some modifications). This result will be presented in a future work.

The remainder of this paper is organized as follows. In Section II we present the code design and proof of the asymptotic optimality of the construction for compressing the discrete nonredundant source with a Hamming distortion criterion. In Section III we propose a suboptimal algorithm for encoding and present empirical results in Section IV, which suggest that the performance of the code and encoding/decoding algorithms is very close to the theoretical limits.

II. CODE CONSTRUCTION AND ANALYSIS FOR THE DISCRETE MEMORYLESS EQUIPROBABLE SOURCE

A. Code construction

We first describe the code design for the binary case and then generalize it to the q -ary alphabet. In the greatest generality a binary (n, k) codebook has blocklength n and 2^k codewords. However if there is no underlying structure to this set (for example a random codebook), then exponential complexity is required to encode/decode. A linear codebook on the other hand is a much restricted set of codewords: all the n -vectors \mathbf{c} , that can be written as

$$\mathbf{c} = \mathbf{G}^T \mathbf{u}, \quad (3)$$

for all possible choices of a k -vector \mathbf{u} , for a given $k \times n$ matrix \mathbf{G} . Note that if \mathbf{u} is not allowed to range over all 2^k choices then the ensuing codebook is in general nonlinear. In fact any codebook (linear or nonlinear) can be described by (3) if \mathbf{u} is allowed to range only over the vectors with unit Hamming weight, and \mathbf{G} has as many rows as codewords, i.e. 2^k .

In this paper we propose a class of nonlinear codebooks that has some convenient structure by letting \mathbf{u} range over the k -vectors with a given Hamming weight $\lceil k\omega \rceil$, where $0 < \omega < 1$. Further, we let

$$k = \lceil n \log_2 n \rceil, \quad (4)$$

and

$$\omega = \frac{R}{\log_2 n \log_2 \log_2 n}. \quad (5)$$

We denote $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M\}$ as the binary k -vectors of Hamming weight $\lceil k\omega \rceil$ in lexicographic order. The codebook is given by $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_M\}$ where $\mathbf{c}_i = \mathbf{G}^T \mathbf{u}_i$. The number of codewords in the codebook is equal to

$$M = \binom{k}{\lceil k\omega \rceil}.$$

Lemma 1. *The asymptotic rate of the code converges to*

$$\lim_{n \rightarrow \infty} \frac{\log_2 M}{n} = R. \quad (6)$$

with the choice of parameters in (4) and (5).

A convenient low density choice of the $k \times n$ matrix \mathbf{G} is by i.i.d. generation of its coefficients where

$$P[\mathbf{G}_{ij} = 1] = \frac{\log_2^2 n}{n}. \quad (7)$$

For the q -ary case the same scheme holds with \mathbf{u} as a binary k -vector, (3) computed with the summation in the group $\{0, 1, \dots, q-1\}$ and (7) generalized to

$$P[\mathbf{G}_{ij} = 0] = 1 - \frac{\log_2^2 n}{n},$$

and

$$P[\mathbf{G}_{ij} = l] = \frac{\log_2^2 n}{n(q-1)}$$

for $l \neq 0$.

B. Code analysis

We now turn to the analysis of the code introduced in Section II-A. We show that with high probability the lossy compressor described in Section II-A asymptotically attains the rate distortion function of the discrete nonredundant source with Hamming distortion. More formally we show the following result.

Theorem 1. *Construct a codebook as outlined in Section II-A with a blocklength n and $R = R(d) + \epsilon$ where $R(d)$ is the rate distortion function for the equiprobable q -ary source with Hamming distortion and $\epsilon > 0$ is arbitrary. Let \mathbf{s} be an arbitrary source realization. If \mathbf{c}_s is the nearest codeword in Hamming distance to the given source then*

$$\lim_{n \rightarrow \infty} P[w_H(\mathbf{s} - \mathbf{c}_s) > nd] = 0.$$

Proof. Pick a random codebook as outlined in Section II-A. Label the codewords as \mathbf{c}_i , $i \in \{1, 2, \dots, M\}$. Denote

$$L_i = \mathbf{1}\{w_H(\mathbf{s} - \mathbf{c}_i) \leq nd\}.$$

and

$$Z = \sum_{i=1}^M L_i.$$

The event $Z > 0$, is equivalent to the event that at least one codeword in the codebook is within Hamming distance d from the given source. Thus to prove the theorem it is sufficient to show that

$$\lim_{n \rightarrow \infty} P[Z > 0] = 1. \quad (8)$$

However (as shown later) using martingale arguments, it is enough to show that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P[Z > 0] = 0, \quad (9)$$

to claim (8). Therefore we will first show (9) and later prove that (9) \implies (8), to complete the proof of Theorem 1 (We use second moment bounds on $P[Z > 0]$ and martingale arguments from [14]). We structure the proof in a sequence of intermediate lemmas. Using the Cauchy-Schwarz inequality we obtain the following lower bound on $P[Z > 0]$, in terms of the first and second moments of the nonnegative random variable Z .

$$P[Z > 0] \geq \frac{E^2[Z]}{E[Z^2]}. \quad (10)$$

To compute $E[Z^2]$, we express it as

Lemma 2.

$$E[Z^2] = E[Z] \left(\sum_{j=1}^M P[L_j = 1 | L_1 = 1] \right). \quad (11)$$

To compute a lower bound on $P[Z > 0]$ we need lower/upper bounds on $E[Z]$ and $\sum_{j=1}^M P[L_j = 1 | L_1 = 1]$ respectively. To that end we present the statistics of the codeword \mathbf{c}_i and the joint statistics of $\mathbf{c}_i, \mathbf{c}_j$ in the following two lemmas.

Lemma 3. *For every $i \in \{1, 2, \dots, M\}$ and $n = 1, 2, \dots$ the q -ary bits $(\mathbf{c}_i[1], \mathbf{c}_i[2], \dots, \mathbf{c}_i[n])$, of the codeword \mathbf{c}_i are independent identically distributed. If $i \neq j$ and A_i, A_j are disjoint subsets of $\{1, 2, \dots, n\}$, then $(\mathbf{c}_i[l], l \in A_i)$ and $(\mathbf{c}_j[l], l \in A_j)$ are independent. Furthermore*

$$\lim_{n \rightarrow \infty} P[\mathbf{c}_i[m] = l] = 1/q,$$

for all $l \in \{0, 1, \dots, q-1\}$.

Proof of Lemma 3. The q -ary bit $\mathbf{c}_i[m] = l$, for $l \neq 0$ if and only if the $\lceil k\omega \rceil$ positions corresponding to the ones in \mathbf{u}_i select some nonzero positions in \mathbf{G} , which sum up to l (modulo q). This event is independent, identically distributed for different m , because the coefficients of \mathbf{G}^T are independent identically distributed. Thus if $A_i \cap A_j = \phi$ then $(\mathbf{c}_i[l], l \in A_i)$ and $(\mathbf{c}_j[l], l \in A_j)$ are independent and the bits $(\mathbf{c}_i[1], \mathbf{c}_i[2], \dots, \mathbf{c}_i[n])$ are independent identically distributed. Furthermore for $l \neq 0$,

$$P[\mathbf{c}_i(m) = l] = \frac{1}{q} \left[1 - \left(1 - \frac{\log^2 n}{n} \right)^{\lceil k\omega \rceil} \right] \quad (12)$$

$$= \frac{1}{q} \left[1 - e^{-\frac{R \log^2 n}{\log \log n}} \right] + o(1) \quad (13)$$

$$\rightarrow \frac{1}{q}. \quad (14)$$

The event $[\mathbf{c}_i(m) = 0]$ is the union of two disjoint events. In the first case $\lceil k\omega \rceil$ ones in \mathbf{u} do not select any nonzero entry. Alternatively they select nonzero entries which sum up to zero. Therefore

$$P[\mathbf{c}_i(m) = 0] = \left[1 - \frac{\log^2 n}{n} \right]^{\lceil k\omega \rceil} + \frac{1}{q} \left[1 - \left(1 - \frac{\log^2 n}{n} \right)^{\lceil k\omega \rceil} \right] \quad (15)$$

$$= e^{-\frac{R \log^2 n}{\log \log n}} + \frac{1}{q} \left[1 - e^{-\frac{R \log^2 n}{\log \log n}} \right] + o(1) \quad (16)$$

$$\rightarrow \frac{1}{q}. \quad (17)$$

Thus

$$\lim_{n \rightarrow \infty} P[\mathbf{c}_i[m] = l] = 1/q \quad \forall l \in \{0, 1, \dots, q-1\}. \quad (18)$$

□

Lemma 4. *If $w_H(\mathbf{u}_i, \mathbf{u}_j) > \frac{\lceil k\omega \rceil}{\log_2 n}$, then the codewords $\mathbf{c}_i, \mathbf{c}_j$ are independent as $n \rightarrow \infty$.*

Proof of Lemma 4. According to Lemma 3, we only need to show that for every $m \in \{1, 2, \dots, n\}$

$$\lim_{n \rightarrow \infty} P[\mathbf{c}_i[m] = a | \mathbf{c}_j[m] = b] = 1/q, \quad (19)$$

where $a, b \in \{0, 1, \dots, q-1\}$. Let $w_H(\mathbf{u}_i, \mathbf{u}_j) = 2f \lceil k\omega \rceil$. Define $\mathbf{u}'_i = \mathbf{w}_{ij}$, $\mathbf{u}'_j = \mathbf{w}_{ji}$ where

$$\mathbf{w}_{ij}[m] = \mathbf{u}_i[m] \cdot \bar{\mathbf{u}}_j[m], \quad (20)$$

and

$$\mathbf{u}'_{ij}[m] = \mathbf{u}_i[m] \cdot \mathbf{u}_j[m]. \quad (21)$$

Where \cdot denotes the logical AND operation and \bar{a} denotes the complement of a . It is easy to see that \mathbf{u}'_i , \mathbf{u}'_j and \mathbf{u}'_{ij} are non-overlapping in the sense that

$$\mathbf{u}'_i \cdot \mathbf{u}'_j = \mathbf{u}'_i \cdot \mathbf{u}'_{ij} = \mathbf{u}'_j \cdot \mathbf{u}'_{ij} = \mathbf{0}.$$

Further

$$\mathbf{u}_i[m] = \mathbf{u}'_i[m] \oplus \mathbf{u}'_{ij}[m]$$

and

$$\mathbf{u}_j[m] = \mathbf{u}'_j[m] \oplus \mathbf{u}'_{ij}[m].$$

Where \oplus denotes the logical XOR operation. Also

$$w_H[\mathbf{u}'_i] = w_H[\mathbf{u}'_j] = f[k\omega],$$

and

$$w_H[\mathbf{u}'_{ij}] = (1 - f)[k\omega].$$

Denote $\mathbf{G}^T \mathbf{u}'_i$, $\mathbf{G}^T \mathbf{u}'_j$ and $\mathbf{G}^T \mathbf{u}'_{ij}$ as \mathbf{c}'_i , \mathbf{c}'_j and \mathbf{c}'_{ij} respectively. These vectors are mutually independent since \mathbf{u}'_i , \mathbf{u}'_j and \mathbf{u}'_{ij} are nonoverlapping. Further, $P[\mathbf{c}'_i[m] = l] \ l \neq 0$ satisfies (see (12)- (13) substituting $[k\omega]$ by $f[k\omega]$)

$$\begin{aligned} P[\mathbf{c}'_i(m) = l] &= \frac{1}{q} \left[1 - \left(1 - \frac{\log^2 n}{n} \right)^{f[k\omega]} \right] \\ &= \frac{1}{q} \left[1 - e^{-\frac{Rf \log^2 n}{\log \log n}} \right] + o(1) \\ &\rightarrow \frac{1}{q}. \end{aligned}$$

Similarly $P[\mathbf{c}'_i[m] = 0]$ satisfies (15) substituting $[k\omega]$ by $f[k\omega]$)

$$\begin{aligned} P[\mathbf{c}'_i(m) = 0] &= \left(1 - \frac{\log^2 n}{n} \right)^{f[k\omega]} + \frac{1}{q} \left[1 - \left(1 - \frac{\log^2 n}{n} \right)^{f[k\omega]} \right] \\ &= e^{-\frac{Rf \log^2 n}{\log \log n}} + \frac{1}{q} \left[1 - e^{-\frac{Rf \log^2 n}{\log \log n}} \right] + o(1) \\ &\rightarrow \frac{1}{q}. \end{aligned}$$

and analogously for j . If V, X_1, X_2 are independent random variables over $\{0, 1, \dots, q-1\}$ with

$$\lim_{n \rightarrow \infty} P[X_i = a] = 1/q \quad \forall a \in \{0, 1, \dots, q-1\}$$

then

$$\lim_{n \rightarrow \infty} P[V + X_1 = a | V + X_2 = b] = 1/q.$$

Where addition is modulo q . Identifying $X_1 = \mathbf{c}'_i[m]$, $X_2 = \mathbf{c}'_j[m]$ and $V = \mathbf{c}'_{ij}[m]$ we get (19). \square

Now returning to the proof of Theorem 1. Let

$$S_l = \{j : w_H(\mathbf{u}_j, \mathbf{u}_1) \leq \frac{[M\omega]}{\log_2 n}\} \quad (22)$$

and

$$S_u = \{j : w_H(\mathbf{u}_j, \mathbf{u}_1) > \frac{[M\omega]}{\log_2 n}\}. \quad (23)$$

In order to compute the quantity in the right hand side of (11) we write

$$\begin{aligned} \sum_{j=1}^M P[L_j = 1 | L_1 = 1] &= \sum_{S_l} P[L_j = 1 | L_1 = 1] \\ &+ \sum_{S_u} P[L_j = 1 | L_1 = 1]. \end{aligned} \quad (24)$$

The first term in (24) is less than or equal to $|S_l|$, which grows sub-exponentially with n as can be shown through a simple counting argument. More formally:

Lemma 5. *Let*

$$S_l = \{j : w_H(\mathbf{u}_j, \mathbf{u}_1) \leq \frac{[M\omega]}{\log_2 n}\},$$

then

$$\lim_{n \rightarrow \infty} \frac{\log |S_l|}{n} = 0. \quad (25)$$

We now compute the exponential rate of decay of $P[L_i = 1]$.

Lemma 6. *Let $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_M\}$ be a random codebook as chosen in Section II-A. For any $\mathbf{s} \in \{0, 1, \dots, q-1\}^n$ let*

$$L_i = \mathbf{1}\{w_H(\mathbf{s} - \mathbf{c}_i) \leq nd\},$$

then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \frac{1}{P[L_i = 1]} = R(d),$$

$\forall i \in \{1, 2, \dots, M\}$, where $R(d)$ is given by (2).

Proof of Lemma 6. From Lemma 3, $\mathbf{c}_i[m]$, $m \in \{1, 2, \dots, n\}$ are independent identically distributed with

$$P[\mathbf{c}_i[m] = a] = p_n.$$

for $a \neq 0$ $n = 1, 2, \dots$, such that $p_n < 1/q$ and

$$\lim_{n \rightarrow \infty} p_n = 1/q.$$

Let $\mathbf{0}, \bar{\mathbf{0}}$ and \mathbf{s} denote the all zero, an arbitrary sequence in $\{1, 2, \dots, q-1\}^n$ and an arbitrary sequence in $\{0, 1, \dots, q-1\}^n$. Since a position in \mathbf{c}_i is more likely to be 0 than any other symbol in the q -ary alphabet we have

$$\begin{aligned} P[w_H(\bar{\mathbf{0}} - \mathbf{c}) \leq nd] &\leq P[w_H(\mathbf{s} - \mathbf{c}) \leq nd] \\ &\leq P[w_H(\mathbf{0} - \mathbf{c}) \leq nd]. \end{aligned} \quad (26)$$

Applying Sanov's theorem on (26) for $n \in \{1, 2, \dots\}$

$$\begin{aligned} \frac{1}{n^2} 2^{-nD(d||1-p_n)} &\leq P[w_H(\mathbf{s} - \mathbf{c}) \leq nd] \\ &\leq 2^{-nD(d||p_n(q-1))} \end{aligned}$$

Therefore

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \frac{1}{P[L_i = 1]} &= D(d||1-1/q) \\ &= \log_2(q) - d \log_2(q-1) - h(d). \end{aligned}$$

\square

Using Lemma 6 and Lemma 1 we have the following result.

Lemma 7. Let $Z = \sum_i L_i$ then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 E[Z] = R - R(d), \quad (27)$$

with $R(d)$ defined in (2).

Now using Lemmas 4 and 6 we get the following result.

Lemma 8. If $R = R(d) + \epsilon$ for an arbitrary $\epsilon > 0$, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \left(\sum_{j=1}^M P[L_j = 1 | L_1 = 1] \right) \leq \epsilon.$$

Proof of Lemma 8. For arbitrary $a_n, b_n > 0$

$$\lim_{n \rightarrow \infty} \frac{\log(a_n + b_n)}{n} = \max \left[\lim_{n \rightarrow \infty} \frac{\log a_n}{n}, \lim_{n \rightarrow \infty} \frac{\log b_n}{n} \right]. \quad (28)$$

Let S_l and S_u be as defined in (24). We have

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \sum_{j=1}^M P[L_j = 1 | L_1 = 1] \\ &= \max_{S \in \{S_l, S_u\}} \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \sum_{j \in S} P[L_j = 1 | L_1 = 1] \\ &= \left[\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \sum_{j \in S_u} P[L_j = 1 | L_1 = 1] \right]^+ \end{aligned} \quad (29)$$

where we get (29) using Lemma 5 and (28). It is easy to verify that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \sum_{j \in S_u} P[L_j = 1 | L_1 = 1] \leq \epsilon, \quad (30)$$

using Lemmas 1, 4 and 6. Combining (29) and (30)

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 \left(\sum_{j=1}^M P[L_j = 1 | L_1 = 1] \right) \leq \epsilon.$$

□

The following result, obtained by using (10) and Lemmas 7 and 8, shows that $P[Z > 0]$, does not decay exponentially in n .

Lemma 9. Let $R = 1 - h(d) + \epsilon$, where $\epsilon > 0$ is arbitrary. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 P[Z > 0] = 0.$$

We can now prove Theorem 1. We will also use the following auxiliary bound.

Lemma 10 ([15]). For a martingale B_1, B_2, \dots, B_n if

$$|B_i - B_{i-1}| < t \quad \forall \quad i \in \{2, 3, \dots, n\},$$

where $t > 0$ is a constant then

$$P[|B_n - B_0| > n\epsilon] < 2e^{-(n/2t^2)\epsilon^2}.$$

Fix a source realization \mathbf{s} . Let

$$f(\mathbf{G}) = \min_j (w_H(\mathbf{c}_j, \mathbf{s})),$$

Where $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_M$ are the codewords. Define a martingale B_i $i \in \{1, 2, \dots, n\}$ in the following manner

$$B_i = E[f(\mathbf{G}^T(X_1, X_2, \dots, X_n)) | X_1, X_2, \dots, X_i]. \quad (31)$$

Where X_1, X_2, \dots, X_n are the rows of the \mathbf{G}^T matrix. Therefore B_0/n is the distortion between the source realization \mathbf{s} and the best codeword present in the code, averaged over all the codebooks. While B_n/n is the distortion between the source realization and the closest codeword in a particular codebook, because all the rows X_1, X_2, \dots, X_n are revealed. $|B_{i+1} - B_i|$ is bounded by 1. Further

$$[Z > 0] \Leftrightarrow [B_n/n \leq d].$$

Suppose

$$\limsup_{n \rightarrow \infty} B_0/n = d + \epsilon',$$

for any $\epsilon' > 0$ then there exists a convergent subsequence of $1, 2, \dots$ such that for this subsequence

$$\lim_{n \rightarrow \infty} B_0/n = d + \epsilon',$$

along this subsequence we have from Lemma 10

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P[B_n/n \leq d] = \lim_{n \rightarrow \infty} \frac{1}{n} \log P[Z > 0] \leq -\epsilon'^2/2.$$

Which contradicts Lemma 9. Hence $\limsup_{n \rightarrow \infty} [B_0/n] \leq d$, which implies that

$$\lim_{n \rightarrow \infty} P[B_n/n > d] = 0, \quad (32)$$

through an application of Lemma 10. Therefore the distortion between any arbitrary source realization \mathbf{s} and the closest codeword in a codebook constructed randomly as given Section II-A, is less than d almost surely as $n \rightarrow \infty$. □

III. SUBOPTIMAL ALGORITHMS FOR COMPRESSION

In this section we describe a suboptimal algorithm to encode a source using the codebook described in Section II. This algorithm attempts to locate a codeword in the codebook which is closest in Hamming distance to the given source. The compressor works as follows. Recall that the codebook is specified by the $k \times n$, matrix \mathbf{G} and $0 < \omega < 1$. It consists of all the codewords \mathbf{c} of length n that can be written as

$$\mathbf{c} = \mathbf{G}^T \mathbf{u},$$

where \mathbf{u} has length k and

$$w_H(\mathbf{u}) = \lceil k\omega \rceil.$$

The algorithm attempts to find a good approximation to the source string \mathbf{s} of length n among the codewords in an iterative manner. At each step in the iteration we select a string of length k , \mathbf{u}_{i+1} by flipping one and only one bit in \mathbf{u}_i . The algorithm starts with $\mathbf{u}_0 = [0, 0, \dots, 0]$ and ends with \mathbf{u}_t such that $w_H(\mathbf{u}_t) = \lceil k\omega \rceil$. Let $\Delta_i = \mathbf{s} - \mathbf{G}^T \mathbf{u}_i$.

The choice of the bit to flip in \mathbf{u}_i is such that $w_H(\Delta_{i+1})$ is minimized. To that end we perform an exhaustive search for all columns of \mathbf{G}^T enumerated as \mathbf{g}_j , $j \in \{1, 2, \dots, k\}$, computing $\Delta_i - \mathbf{g}_j$, and selecting the index j that leads to the lowest Hamming distance. This procedure is repeated till $w_H(\Delta_i) > w_H(\Delta_{i+1})$. If $w_H(\Delta_i) \leq w_H(\Delta_{i+1})$, then if $w_H(\mathbf{u}_i) < \lceil k\omega \rceil$, the algorithm is now constrained to flip only bits which have a value of zero in \mathbf{u}_i which lead to minimum $w_H(\Delta_{i+1})$, and vice versa for the case when $w_H(\mathbf{u}_i) > \lceil k\omega \rceil$. We halt when $w_H(\mathbf{u}_t) = \lceil k\omega \rceil$. It is immaterial how ties are broken by the compressor. At the final configuration of \mathbf{u}_t , the encoder then stores the value of \mathbf{u}_t in the form of an index, using an enumerative encoding scheme [16]. The decoder then uses this index to recover \mathbf{u}_t , and outputs $\mathbf{G}^T \mathbf{u}_t$. The complexity of the algorithm is $O(n^2 \log_2^3(n))$, compared to various message passing based approaches such as survey propagation [12] and its variants [13], which incur a complexity of $O(n^2)$, with inferior compression efficiency.

IV. EXPERIMENTS

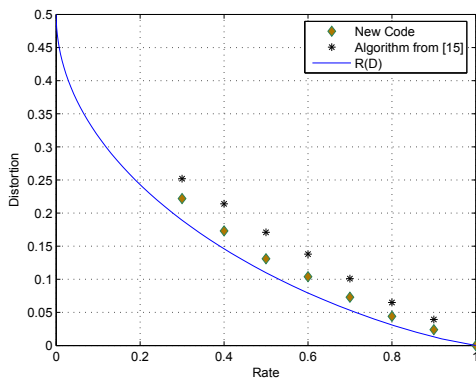


Fig. 1. Empirical performance of our code/algorithms compared with the message passing heuristic from [13] for the binary symmetric source with blocklength 400.

We show empirical results obtained with the codes described in Section II, and the encoding algorithm in Section III. For each rate we fix a randomly generated codebook and average the distortion obtained for compressing a random source (for 1000 iterations). In Figure 1 we compare our algorithm with a message passing algorithm that had the best performance in the literature for compressing the binary symmetric source using LDGM codes (an implementation of the algorithm in [13]). The algorithms are compared at a short blocklength ($n=400$) because at larger blocklengths both algorithms perform close to optimal and the differences are harder to discern. Figure 1 demonstrates that our code design and algorithms are competitive with the state of the art in lossy data compression. In Figure 2 we show results obtained from compressing the nonredundant 16-ary source with blocklength 1024 and Hamming distortion criterion. This figure shows excellent performance of our codes for moderate blocklengths for compressing a general (nonbinary) source.

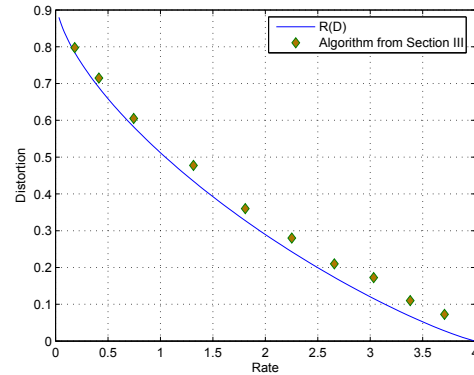


Fig. 2. Empirical performance of the codes for the 16-ary source for blocklength 1024.

REFERENCES

- [1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Convention*, vol. 7, pp. 142–163, 1959.
- [2] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, pp. 460–473, July 1972.
- [3] T. Berger, *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall Eaglewood Cliffs NJ, 1971.
- [4] I. Csiszar and J. Korner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*. New York Academic, 1981.
- [5] T. C. Anчета, "Syndrome source coding and its universal generalizations," *IEEE Transactions on Information Theory*, vol. 22, pp. 423–428, March 1976.
- [6] G. Caire, S. Shamai, and S. Verdú, "Lossless data compression with error correcting codes," *Advances in Network Information Theory, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society*, vol. 66, pp. 263–284, 2004.
- [7] G. Caire, S. Shamai, A. Shokrollahi, and S. Verdú, "Fountain codes for lossless data compression," *Algebraic Coding Theory and Information Theory, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 68, pp. 1–20, 2006.
- [8] T. J. Goblick, "Coding for a discrete information source with a distortion measure," Ph.D. dissertation, Department of Electrical Engineering MIT, Massachusetts, 1962.
- [9] Y. Matsunaga and H. Yamamoto, "A coding theorem for lossy data compression by LDPC codes," *IEEE Transactions on Information Theory*, vol. 49, September 2003.
- [10] D. J. C. Mackay, "Good error-correcting codes based on very sparse matrices," *IEEE Transactions on Information Theory*, vol. 49, pp. 399–431, March 1999.
- [11] S. Miyake, "Lossy data compression over Z_q by LDPC code," *IEEE International Symposium on Information Theory*, 2006.
- [12] S. Ciliberti, K. Mezard, and R. Zecchina, "Message passing algorithms for non-linear nodes and data compression," *arxiv*. [Online]. Available: <http://arxiv.org/abs/cond-mat/0508723>
- [13] E. Maneva and M. J. Wainwright, "Lossy source encoding via messagepassing and decimation over generalized codewords of LDGM codes," *IEEE International Symposium on Information Theory*, September 2005.
- [14] E. Martinian and M. J. Wainwright, "Low-density codes achieve the rate-distortion bound," *Data Compression Conference*, 2006.
- [15] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. Journal*, vol. 19, pp. 357–367, 1967.
- [16] T. M. Cover, "Enumerative source encoding," *IEEE Transactions on Information Theory*, vol. 10, pp. 460–473, January 1973.