Lower bounds on the rate-distortion function of LDGM codes

A. G. Dimakis[1], M. J. Wainwright[1,2] and K. Ramchandran[1],
[1]Department of Electrical Engineering and Computer Science
[2]Department of Statistics
UC Berkeley, Berkeley, CA.

Abstract— **We analyze the performance of low-density generator matrix (LDGM) codes for lossy source coding. We first develop a generic technique for deriving lower bounds on the effective rate-distortion functions of binary linear codes. This result provides a source coding analog of a classical result due to Gallager for channel coding over the binary symmetric channel. We illustrate this method for the ensemble of check-regular low-density generator matrix (LDGM) codes by deriving an explicit lower bound on its rate-distortion performance as a function of the check degree.**

I. INTRODUCTION

Classical random coding arguments show that random binary linear codes will achieve the rate-distortion bound for lossy compression of a symmetric Bernoulli source. However, such codes are impractical, as it is neither possible to represent them in an compact manner, nor to perform encoding/decoding in an efficient way. It is thus of considerable interest to explore and analyze the use of sparse graphical codes for lossy compression problems. A line of recent work [13], [18], [3], [16] has explored the use of low-density generator matrix (LDGM) codes for lossy compression. The results of this paper provide further insight into the effective rate-distortion function of this class of sparse graph codes.

Past work: One practical approach to lossy compression is via trellis-code quantization (TCQ) [9]. One limitation of trellis-based approaches is the fact that saturating rate-distortion bounds requires increasing the trellis constraint length [17], which incurs exponential complexity (even for the max-product or sum-product message-passing algorithms). Other work [14] shows that it is possible to approach the binary rate-distortion bound using LDPC-like codes, albeit with degrees that grow logarithmically with the blocklength. A parallel line of work has studied the use of low-density generator matrix (LDGM) codes, which correspond to the duals of LDPC codes, for lossy compression problems [13], [18], [3], [16]. Focusing on binary erasure quantization (a special compression problem dual to binary erasure channel coding), Martinian and Yedidia [13] proved that LDGM codes combined with modified message-passing can saturate the associated rate-distortion bound. Various researchers have used techniques from statistical physics, including the cavity method and replica methods, to provide non-rigorous analyses of LDGM performance for lossy compression of binary sources [2], [3], [16]. In the limit of zero-distortion, this

analysis has been made rigorous in a sequence of papers [5], [15], [4], [6]. Moreover, our own recent work [11], [10] provides rigorous *upper bounds* on the effective rate-distortion function of various classes of LDGM codes. In terms of practical algorithms for lossy binary compression, researchers have explored variants of the sum-product algorithm [16] or survey propagation algorithms [2], [18] for quantizing binary sources.

Our contributions: Previous analysis of LDGM rate-distortion [11], [12] was based on the first and second-moment methods from probabilistic combinatorics [1]. Whereas the second moment provides a non-trivial upper bound on the effective rate-distortion function, the first moment method yields a well-known statement—namely, the Shannon bound, which is far from sharp for these sparse graph codes. The primary contribution of this paper is the development of a technique for generating *sharper lower bounds* on the effective rate-distortion function of sparse graph codes. Our approach can be understood as a source coding analog of Gallager's [7] classical result on the effective capacity of bounded degree LDPC codes for channel coding. We illustrate our approach in application to the check-regular ensemble of LDGM codes, establishing how its effective rate-distortion performance remains bounded away from the Shannon limit for any finite check degree.

Theorem 1. *Let $\mathbb{C}_{CR}$ be an LDGM code randomly drawn from the check-regular ensemble with degree d_c, and suppose that source encoding/decoding are performed optimally. With high probability, the LDGM code $\mathbb{C}_{CR}$ can only achieve those rate-distortion pairs (R, D) that satisfy the bound*

$$R \left[1 - \exp(-\frac{(1-D)d_c}{R}) \right] \ \geq \ 1 - H(D). \quad (1)$$

For any finite degree d_c, the minimal rate R satisfying the relation (1) *is strictly bounded away from the Shannon rate.*

We note that the bound (1) is strictly tighter than the obvious bound for all $D > 0$, based on counting isolated information bits in the check-regular ensemble. However, we also know that the bound is not sharp, and could be refined through more careful analysis.

The remainder of this paper is organized as follows. Section II contains basic background material and definitions for source coding, factor graphs, and low-density generator matrix codes. Section III is devoted to a number of basic results,

applicable to any binary linear code. Our analysis of the check-regular ensemble of LDGM codes is given in Section IV.

II. BACKGROUND

A. Binary codes and source coding

In abstract terms, a binary linear code $\mathbb{C}$ of block length n consists of a linear subspace of $\{0,1\}^n$. One concrete representation is as the range space of a given generator matrix $G \in \{0,1\}^{n \times m}$, as follows:

$$\mathbb{C} = \{x \in \{0,1\}^n \mid x = Gz \quad \text{for some } z \in \{0,1\}^m\} \quad (2)$$

The code $\mathbb{C}$ consists of at most $2^m = 2^{nR}$ codewords, where $R = \frac{m}{n}$ is the code rate.

In the binary lossy source coding problem, the encoder observes a symmetric Bernoulli source sequence $S \in \{0,1\}^n$, with each element S_i drawn in an independent and identically distributed (i.i.d.) manner from a Bernoulli distribution with parameter $p = \frac{1}{2}$. The idea is to compress the source by representing each source sequence S by some codeword $x \in \mathbb{C}$. When using a code in generator matrix form, one thinks of mapping each source sequence to some codeword $x \in \mathbb{C}$ from a code containing $2^m = 2^{nR}$ elements, say indexed by the binary sequences $z \in \{0,1\}^m$. The source decoding map $x \mapsto \widehat{S}(x)$ associates a source reconstruction $\widehat{S}(x)$ with each codeword $x \in \mathbb{C}$. The quality of the reconstruction can be measured in terms of the Hamming distortion $d(S, \widehat{S}) = \sum_{i=1}^n |S_i - \widehat{S}_i| = \|S - \widehat{S}\|_1$. With this set-up, the source encoding problem is to find the codeword with minimal distortion—namely, the optimal encoding $\widehat{x}_{ML} := \arg\min_{x \in \mathbb{C}} d(\widehat{S}(x), S)$. Classical rate-distortion theory dictates that, for the binary symmetric source, the optimal trade-off between the compression rate R and the best achievable average distortion $D = \mathbb{E}[d(\widehat{S}, S)]$ is given by the function $R(D) = 1 - h(D)$, where h is the binary entropy function.

B. Factor graphs and LDGM codes

Given a binary linear code $\mathbb{C}$, specified by generator matrix G, the code structure can be captured by a bipartite graph, in which square nodes (■) represent the checks attached to the code output bits x_i (or rows of G), and circular nodes (○) represent the information bits (or columns of G). For instance, Fig. 1 shows the factor graph for a rate $R = \frac{3}{4}$ code in generator matrix form, with $n = 12$ checks (each associated with a unique source bit, top of diagram) connected to a total of $m = 9$ information bits (bottom of diagram). The edges in this graph correspond to 1's in generator matrix matrix, and reveal the subset of bits to which each information bit contributes. The degrees of the check (respectively) variable nodes in the factor graph are $d_c = 3$ and $d_v = 4$ respectively, so that the associated generator matrix G has 3 ones in each row, and 4 ones in each column. When the generator matrix is sparse, then the resulting code is known as a low-density
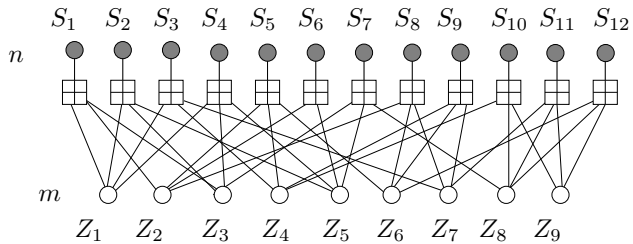

Fig. 1. Factor graph representation of an LDGM code with $n = 12$ source bits, $m = 9$ checks, and overall rate $R = \frac{3}{4}$.

generator matrix (LDGM) code. As an example, the *check-regular LDGM ensemble* that we analyze later is formed by fixing a check degree d_c, and having each of the n checks connect to d_c of the information bits uniformly at random (with replacement).

III. BASIC RESULTS

In this section, we develop a number of basic results, applicable to any binary linear code, which underlie our analysis of LDGM codes, to follow in Section IV.

A. Gallager-style bound

A classical approach to analyzing the performance of LDPC codes for channel coding, originally due to Gallager [7], is based decomposing the mutual information between the channel input and output in two ways, and linking these terms to the code rate and error probability. Here we develop an analogous approach for analyzing the rate-distortion performance of binary linear codes.

Given the generator matrix $G \in \{0,1\}^{n \times m}$ of a binary linear code, codewords in $\mathbb{C}$ are all of the form Gz, where $z \in \{0,1\}^m$ is a sequence of *information bits*. Given some source sequence $S \in \{0,1\}^n$ of symmetric Bernoulli random variables, suppose that we quantize the source using the code $\mathbb{C}$. We use $\widehat{S} \in \{0,1\}^n$ to denote the codeword to which the random sequence S is quantized; for now, we leave the precise nature of the encoder mapping $S \mapsto \widehat{S}$ unspecified. We denote by $\widehat{Z} \in \{0,1\}^m$ a sequence chosen uniformly at random from all information sequences that generate $\widehat{S}$ via the relation $\widehat{S} = G\widehat{Z}$. Assume that this source encoder applied to $\mathbb{C}$ achieves average distortion D, in that

$$\frac{1}{n}\mathbb{E}[\|S \oplus \widehat{S}\|_1] \leq D,$$

where the expectation $\mathbb{E}$ is taken over the random source sequence, as well as any possible randomness in the mapping $S \mapsto \widehat{S}$. Under this assumption, we have

Lemma 1. *For any code $\mathbb{C}$ with blocklength n and rate R with an encoder achieving average distortion D, the rate and distortion are linked via the bound:*

$$R \geq 1 - H(D) + \frac{1}{n}H(\widehat{Z} \mid S). \quad (3)$$

Proof: The mutual information between S and $\widehat{S}$ is given by $I(S;\widehat{S}) = H(S) - H(S \mid \widehat{S})$. Observe that $H(S) = n$ since S is an i.i.d. sequence of symmetric Bernoulli random variables. Moreover, since the average distortion is upper bounded by D, we have $H(S \mid \widehat{S}) \leq nH(D)$. To see this, consider a fixed encoder that maps $s \mapsto \widehat{S}$ with average distortion $\mathbb{E}\|S \oplus \widehat{S}\| = nD$. For each coordinate define $Z_i = S_i \oplus \widehat{S}_i$ and for these Bernoulli random variables let $D_i = \mathbb{P}(Z_i = 1)$. We therefore have $\sum_{i=1}^n \mathbb{E} Z_i = \sum_{i=1}^n D_i = nD$. We can now bound the entropy

$$
\begin{aligned}
H(S \mid \widehat{S}) &\leq \sum_{i=1}^n H(S_i \oplus \widehat{S}_i) \\
&= \sum_{i=1}^n H(D_i) = n \sum_{i=1}^n \frac{1}{n} H(D_i).
\end{aligned}
$$

By exploiting the concavity of entropy and applying Jensen's inequality, we obtain that $\sum_{i=1}^n \frac{1}{n} H(D_i) \leq H(\frac{1}{n}\sum_{i=1}^n D_i) = H(D)$, which shows that $H(S \mid \widehat{S}) \leq nH(D)$. Consequently, we have the lower bound

$$
\frac{1}{n} I(S;\widehat{S}) \geq 1 - H(D). \tag{4}
$$

On the other hand, since $\widehat{S}$ is a function of $\widehat{Z}$, the data processing inequality implies that $I(S;\widehat{S}) \leq I(S;\widehat{Z})$. Moreover, we have

$$
\begin{aligned}
\frac{1}{n} I(S;\widehat{Z}) &= \frac{1}{n} H(\widehat{Z}) - \frac{1}{n} H(\widehat{Z} \mid S) \\
&\leq R - \frac{1}{n} H(\widehat{Z} \mid S), \tag{5}
\end{aligned}
$$

since there are $2^m = 2^{nR}$ different information sequences. Combining the lower bound (4) with the upper bound (5) yields the claim (3). ∎

Remark: If we used the trivial lower bound $\frac{1}{n} H(\widehat{Z} \mid S) \geq 0$, then the bound (3) would reduce to the Shannon rate-distortion bound. Indeed, this bound would be asymptotically tight for a random linear code. For other codes, obtaining more refined statements requires exploiting specific aspects of the code structure.

B. Maximum likelihood and D-ball encoding

Given a code $\mathbb{C}$, the optimal encoder is the so-called maximum likelihood (ML) encoder. Given a source sequence, it computes the set of codewords closest in Hamming distance, and outputs one of them uniformly at random, as $\widehat{S}_{\mathrm{ML}} \in \arg\min_{x \in \mathbb{C}}\{\|x \oplus S\|_1\}$. The associated minimal distortion is a random variable, defined as

$$
\begin{aligned}
d_n(S;\mathbb{C}) &:= \min_{x \in \mathbb{C}}\left\{\frac{1}{n}\|x \oplus S\|_1\right\} \\
&= \frac{1}{n}\|\widehat{S}_{\mathrm{ML}} \oplus S\|_1. \tag{6}
\end{aligned}
$$

This ML encoder is optimal in that its expected distortion $\mathbb{E}[d_n(S;\mathbb{C})]$ is minimized over all encoders.

Despite the optimality of ML encoding, it is more convenient for theoretical purposes to analyze the following D-ball encoder. For any fixed target distortion $D \in (0, \frac{1}{2})$, define the Hamming ball of radius D around the source sequence S as follows:

$$
\mathbb{B}_n(S;D) := \{x \in \{0,1\}^n \mid \|S \oplus x\|_1 \leq Dn\}. \tag{7}
$$

We say that the D-ball encoder *succeeds* if and only if the intersection $\mathbb{B}_n(S;D) \cap \mathbb{C}$ is non-empty, in which case it chooses some $\widehat{S}_{\mathrm{DB}}$ uniformly at random from this intersection. Otherwise, the encoder fails, and we set $\widehat{S}_{\mathrm{DB}}$ equal to some codeword chosen uniformly at random from the code $\mathbb{C}$.

We now claim that the D-ball encoder is asymptotically equivalent to the ML encoder.

Lemma 2. *For any binary linear code, the following two conditions are equivalent:*
(a) *for all $\epsilon > 0$, the probability of success under $(D+\epsilon)$-ball encoding converges to one as $n \to +\infty$.*
(b) *for all $\delta > 0$, we have $\mathbb{E}[d_n(S;\mathbb{C})] \leq D + \delta$ for all suitably large blocklengths n.*

Proof: We first show that (a) implies (b). Given any fixed $\delta > 0$, set $\epsilon = \delta/2$ in part (a), and consider the associated $(D+\frac{\delta}{2})$-ball encoder. Setting $p_n = \mathbb{P}[(D + \delta/2)$-ball success], we have

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}[\|\widehat{S}_{\mathrm{DB}} \oplus S\|_1] &\leq (D + \frac{\delta}{2})p_n + (1 - p_n)\frac{1}{2} \\
&\leq D + \frac{1}{2}[1 - p_n + p_n\delta] \\
&\leq D + \delta,
\end{aligned}
$$

where the final inequality follows if we can ensure that $p_n \geq \frac{1-2\delta}{1-\delta}$. Since $p_n \to 1$ by assumption, this condition can be met by choosing n sufficiently large. Finally, since ML encoding yields the minimal average distortion, we have

$$
\mathbb{E}[d_n(S;\mathbb{C})] \leq \frac{1}{n}\mathbb{E}[\|\widehat{S}_{\mathrm{DB}} \oplus S\|_1] \leq D + \delta,
$$

which is the claim (b).

We now prove that NOT (a) implies NOT (b). Suppose that for some $\epsilon > 0$, the encoding success probability $p_n = \mathbb{P}[(D + \epsilon)$-ball success] does not converge to 1. Then $\liminf p_n < 1$, so that by taking subsequences if necessary, we may assume that for all sufficiently large n, the failure probability satisfies $1 - p_n \geq \nu$ for some $\nu > 0$. Since the $(D + \epsilon)$-ball encoder can fail only if there are no codewords within normalized distance $(D + \epsilon)$ of the source sequence, this statement implies $\mathbb{P}[d_n(S;\mathbb{C}) > D + \epsilon] \geq \nu$.

Next, we claim that the ML distortion $d_n(S;\mathbb{C})$ is concentrated around its expected value. It is not hard to see that changing one bit S_i cannot change $d_n(S;\mathbb{C})$ by more than $1/n$. Therefore, by applying the Azuma-Hoeffding inequality [8] to the martingale sequence formed by the c-Lipschitz function $d_n(S;\mathbb{C})$ for $c = 1/n$ yields the concentration:

$$
\mathbb{P}[|d_n(S;\mathbb{C}) - \mathbb{E}[d_n(S;\mathbb{C})]| \geq \epsilon] \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right). \tag{8}
$$

Therefore, for any constant $\epsilon > 0$, we have

$$
\lim_{n \to +\infty} \mathbb{P}[|d_n(S;\mathbb{C}) - \mathbb{E}[d_n(S;\mathbb{C})]| \geq \epsilon] = 0. \tag{9}
$$

Using this concentration (9), we see that the bound $\mathbb{P}[d_n(S;\mathbb{C}) > D + \epsilon] \geq \nu$ implies that $\mathbb{E}[d_n(S;\mathbb{C})] \geq D + \epsilon$. (Otherwise, the tail decay would be violated.) Hence, we have established the existence of some $\epsilon > 0$ for which there exists an infinite sequence of blocklengths for which $\mathbb{E}[d_n(S;\mathbb{C})] \geq D + \epsilon$, thus implying NOT (b).

■

C. Bounding the conditional entropy of information bits

Recall from Lemma 1 that lower bounds on the rate-distortion function can be obtained via lower bounds on $\frac{1}{n}H(\widehat{Z} \mid S)$. Since our analysis will involve bounding the conditional entropy of $\widehat{Z}$, recall that we assume that after choosing some sequence $\widehat{S}$, the D-ball encoder then chooses an information sequence $\widehat{Z}$ uniformly at random from all information sequences satisfying $G\widehat{Z} = \widehat{S}$. We then have the following

Lemma 3. *Let U be a random variable in $\mathbb{B}_n(0; D)$ distributed as $U \overset{d}{=} (S \mid \widehat{S} = 0, D\text{-ball success})$. Then the conditional entropy $H(\widehat{Z} \mid S)$ is lower bounded by*

$$p_n\, \mathbb{E}_U\left[\log |\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(U;D)\}|\right], \quad (10)$$

where $p_n := \mathbb{P}[D\text{-ball success}]$.

Proof: By definition, we have

$$H(\widehat{Z} \mid S) := \mathbb{E}_S\left[\sum_z \mathbb{P}(z \mid S) \log \frac{1}{\mathbb{P}(z \mid S)}\right],$$

where $\mathbb{P}(z \mid S)$ denotes the conditional distribution over information sequences $\widehat{Z}$ that the encoder can output for a given source sequence S. First, we claim that by definition of our encoding rule, when D-ball encoding succeeds, the conditional distribution $\mathbb{P}(z \mid S)$ is uniform over the set of all information sequences z such that $Gz \in \mathbb{B}_n(s; D)$. By definition, our encoder chooses uniformly from the set of codewords $\widehat{s} \in \mathbb{B}_n(s; D)$. To ensure that this uniformity is preserved over the information sequences as well, we simply need to verify that each codeword $\widehat{s}$ is generated by the same number—say k—of information sequences. Suppose that we have

$$0 = G\vec{0} = Gz_2 = \ldots = Gz_k,$$

where $(z_2, \ldots, z_k)$ are $(k-1)$ distinct and non-zero information sequences. Let $\widehat{s}$ be any non-zero codeword, say with $\widehat{s} = Gz_0$ for some non-zero information sequence $z_0 \in \{0,1\}^m$. Then the information sequences $(z_0, z_2 \oplus z_0, \ldots, z_k \oplus z_0)$ constitute k distinct generators of $\widehat{s}$.

As a consequence of this uniformity, for each source sequence s such that D-ball encoding succeeds, we have

$$\log \frac{1}{\mathbb{P}(z \mid s)} = \log |\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(s; D)\}|.$$

The structure of a binary linear code is invariant to translations by codewords. Accordingly, we may translate the center of the

ball by the codeword Gz, thus yielding that $\log \frac{1}{\mathbb{P}(z|s)}$ is equal to

$$\log |\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(s \oplus Gz; D)\}|.$$

Taking averages over the z, we have that $\sum_z \mathbb{P}(z \mid s) \log \frac{1}{\mathbb{P}(z|s)}$ is lower bounded by

$$\sum_z \mathbb{P}(z \mid s) \log |\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(s \oplus Gz; D)\}|$$

$$\geq \sum_{\widehat{s}} \mathbb{P}(\widehat{s} \mid s) \log |\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(s \oplus \widehat{s}; D)\}|$$

$$= \mathbb{E}_{\widehat{S}}\left[\log \left|\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(s \oplus \widehat{S}; D)\}\right| \mid S = s\right].$$

Finally, conditioning on the event of D-ball success and taking conditional expectations over S, we have

$$H(\widehat{Z} \mid S) = \mathbb{E}_S\left[\sum_z \mathbb{P}(z \mid S) \log \frac{1}{\mathbb{P}(z \mid S)}\right],$$

which is in turn lower bounded by

$$p_n \mathbb{E}_{S,\widehat{S}}\left[\log \left|\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(S \oplus \widehat{S}; D)\}\right|\right],$$

where this final expectation is conditioned on the success of the D-ball encoder. To complete the proof, we note that by linearity of the code, for any codeword $\widehat{S}$, the cardinality $\left|\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(S \oplus \widehat{S}; D)\}\right|$ is equal to $|\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(S; D)\}|$, so that the conditional expectation over $(S, \widehat{S})$ can be rewritten as a single expectation over U as defined in the lemma statement.

■

Remark: A challenge associated with the analysis of the lower bound (10) concerns the distribution of the random variable U. In general, it is not simply uniform over $\mathbb{B}_n(0; D)$, since the probability of different source sequences $S = s$ given that $\widehat{S} = 0$ varies depending on how many other codewords belong to the ball $\mathbb{B}_n(s; D)$. As a particular example, for the trivial code $\mathbb{C} = \{00, 11\}$ of blocklength $n = 2$ and $Dn = 1$, we have $\mathbb{P}(S = 00 \mid \widehat{S} = 00, D\text{-ball success}) = \frac{1}{2}$, and $\mathbb{P}(S = 01 \mid \widehat{S} = 00, D\text{-ball success})$ and $\mathbb{P}(S = 10 \mid \widehat{S} = 00, D\text{-ball success})$ both equal to $\frac{1}{4}$.

IV. LOWER BOUNDS FOR LDGM CODES

With these basic results, we now turn to lower bounding the rate-distortion functions of LDGM code ensembles.

A. Graph-based certificate

Our goal is to provide a concrete graph-based condition that allows us to lower bound the quantity

$$\mathbb{A}(U; \mathbb{C}) := \log |\{z' \in \{0,1\}^m \mid Gz' \in \mathbb{B}_n(U; D)\}|. \quad (11)$$

Note that $\mathbb{A}(U; \mathbb{C})$ is a random variable, even when the code $\mathbb{C}$ is fixed. In order to analyze this quantity, it is convenient to develop a further lower bound.

For any $U \in \mathbb{B}_n(0; D)$, we let $\mathbb{S}^{\text{fix}}(U)$ be any collection of $(1-D)n$ check bits for which $U_i = 0$. Note that the inclusion $U \in \mathbb{B}_n(0; D)$ guarantees the existence of at least one such
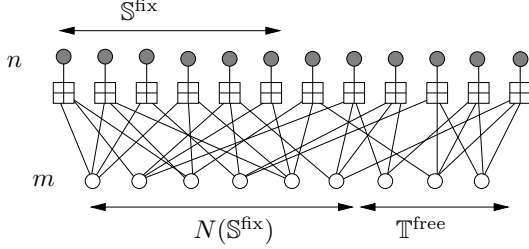
Fig. 2. Tanner Graph of an LDGM code illustrating the fixed checks $\mathbb{S}^{\text{fix}}$, information bit neighbors $N(\mathbb{S}^{\text{fix}})$ of fixed checks, and the free information bits $\mathbb{T}^{\text{free}}$.

subset; otherwise, we simply choose $\mathbb{S}^{\text{fix}}(U)$ uniformly at random from all possible subsets. Associated with $\mathbb{S}^{\text{fix}}(U)$ is a subset of the information bits, namely

$$\mathbb{T}^{\text{free}}(U;\mathbb{C}) := \{j \in \{1,\ldots,m\} \mid j \notin N(\mathbb{S}^{\text{fix}}(U))\}, \quad (12)$$

where for any subset $A \subseteq \{1,\ldots,n\}$, we use $N(A)$ to denote the information bit neighborhood of the variables in A. See Figure 2 for an illustration of these concepts.

The key observation is the following: any information bit $z_i \in \mathbb{T}^{\text{free}}(U;\mathbb{C})$ is effectively free, since we may alter it arbitrarily while still ensuring that $Gz \in \mathbb{B}_n(U;D)$. Therefore, for all codes and $U \in \mathbb{B}_n(0;D)$, we have the lower bound $\mathbb{A}(U;\mathbb{C}) \geq \left|\{\mathbb{T}^{\text{free}}(U;\mathbb{C})\}\right|$. Hence, applying Lemma 3, we can lower bound the conditional entropy as

$$\frac{1}{p_n}H(\widehat{Z} \mid S) \geq \mathbb{E}_U\left[\left|\{\mathbb{T}^{\text{free}}(U;\mathbb{C})\}\right|\right] =: W(\mathbb{C}), \quad (13)$$

where $p_n = \mathbb{P}[D\text{-ball success}]$. In the regime of interest, we have $p_n \to 1$, so that we focus on analyzing the quantity $W(\mathbb{C})$.

B. Analysis over random ensembles

For any fixed code $\mathbb{C}$, the quantity $W(\mathbb{C})$ is deterministic and it provides a *certificate* of the excess rate incurred by using the given code. For a fixed code, however, it is nontrivial to compute the expectation over U defining $W(\mathbb{C})$ in equation (13), since the distribution of U is controlled by the code structure in a non-trivial manner. However, performing the analysis over a random ensemble of LDGM codes allows us to circumvent this problem. Accordingly, our method consists of the following steps:

(a) For a given code ensemble $\mathfrak{C}$, we consider the *random variable* $W(\mathbb{C})$, where $\mathbb{C} \sim \mathfrak{C}$ is a randomly drawn code, and show that the expectation over the code ensemble $\mathbb{E}_{\mathbb{C}}[W(\mathbb{C})]$ scales linearly in blocklength.

(b) Next we show that the random variable $W(\mathbb{C})$ is concentrated around its mean, and use this to establish a lower bound on the random variable $W(\mathbb{C})$ that holds with probability one as the blocklength increases.

Due to space constraints, here we limit our explicit illustration of this approach to the check-regular ensemble.

C. The check-regular ensemble

Recall the check-regular ensemble of LDGM codes, denoted by $\mathfrak{C}_{CR}$. consists of codes with n checks and m information bits, constructed by having each check select d_c bits uniformly at random (and with repetition). The expected number of free bits over the check-regular code ensemble can be easily computed, since the randomization over codes has a symmetrization effect. As a consequence, we may assume without loss of generality that the first $(1-D)n$ elements of the n checks are fixed.

Lemma 4. *The expectation of $W(\mathbb{C}_{CR})$ over the check-regular ensemble is grows linearly in blocklength:*

$$\mathbb{E}[W(\mathbb{C}_{CR})] = m\left(1 - \frac{1}{m}\right)^{(1-D)nd_c}. \quad (14)$$

Further, with probability converging to one, the random variable $W(\mathbb{C}_{CR})$ does not deviate from its expectation by more than $c\sqrt{m\ln m}$ for a suitable constant c, in the sense that

$$\mathbb{P}[|W(\mathbb{C}_{CR}) - \mathbb{E}W(\mathbb{C}_{CR})| \geq c\sqrt{m\ln m}] \leq \frac{2}{m}. \quad (15)$$

Proof: The probability that a particular bit is free (i.e. non-adjacent to $\mathbb{S}^{\text{fix}}(U)$) is simply

$$p_i = \left(1 - \frac{1}{m}\right)^{(1-D)nd_c}, \quad (16)$$

since it is not adjacent to a particular check with probability $(\frac{m-1}{m})^{d_c}$ and that happens for all the $(1-D)R$ fixed checks independently. Therefore, the expected size of $\mathbb{T}^{\text{free}}$ is simply $\mathbb{E}W(\mathbb{C}_{CR}) = mp_i$. Observe that

$$\lim_{m\to\infty} p_i = \exp\left(-\frac{(1-D)d_c}{R}\right), \quad (17)$$

a standard Poisson limit.

Now we need to show that the random variable $W(\mathbb{C}_{CR})$ is concentrated around its mean. Using a Chernoff bound [8], for any $\delta \in [0,1)$,

$$\mathbb{P}\left[|W(\mathbb{C}_{CR}) - \mathbb{E}W(\mathbb{C}_{CR})| \geq \delta\mathbb{E}W(\mathbb{C}_{CR})\right]$$
$$\leq 2\exp\left(-\frac{\delta^2\mathbb{E}W(\mathbb{C}_{CR})}{3}\right).$$

Using $\mathbb{E}W(\mathbb{C}_{CR}) = mp_i$ and by setting $\delta = c\sqrt{\frac{\ln m}{m}}$, $c = \sqrt{3/p_i}$, we find

$$\mathbb{P}[|W(\mathbb{C}_{CR}) - \mathbb{E}W(\mathbb{C}_{CR})| \geq c\sqrt{m\ln m}] \leq \frac{2}{m}. \quad (18)$$

$\blacksquare$

To conclude, we define the critical rate-loss exponent for the check regular ensemble:

$$\gamma_{CR}(m) = \left(\exp[-\frac{(1-D)d_c}{R}] - \frac{\epsilon_1}{m}\right)\left(1 - \frac{2}{m}\right). \quad (19)$$

We therefore obtain the following bound on rate-distortion for any blocklength m

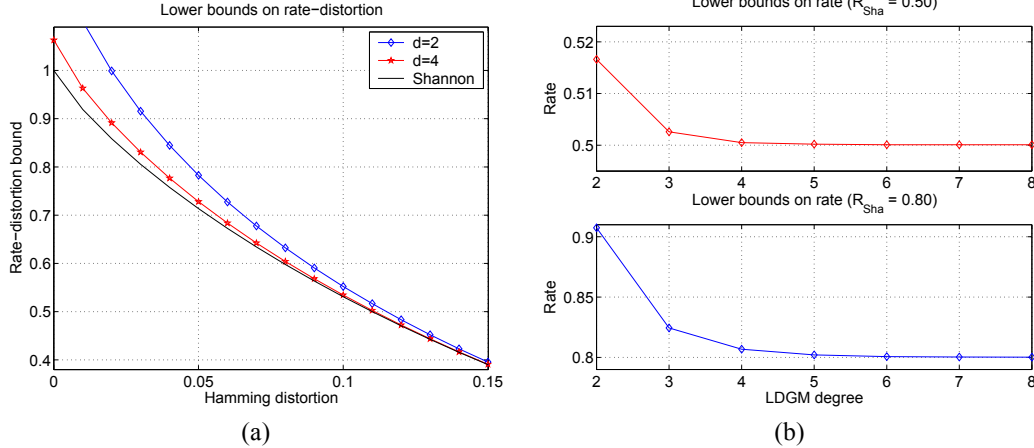$$R\left[1 - \gamma_{CR}(m)\right] \geq (1 - H(D)). \quad (20)$$

Fig. 3. (a) (a) Plots of lower bound (21) on the effective rate-distortion function of the check-regular LDGM ensemble for degrees $d_c = 2$ and $d_c = 4$, as compared to the Shannon limit. (b) Convergence to the Shannon limit as the degree is increased.

Taking limits as $m \to +\infty$, we have $\lim_{m\to\infty} \gamma_{CR}(m) = \exp(-\frac{(1-D)d_c}{R})$. Overall, we conclude that with high probability, a code drawn randomly from the check-regular LDGM ensemble with degree d_c can only obtain rate-distortion pairs (R, D) that satisfy

$$R \left[1 - \exp(-\frac{(1-D)d_c}{R}) \right] \ \geq \ 1 - H(D). \quad (21)$$

V. DISCUSSION

We developed a technique for generating lower bounds on the effective rate-distortion function of sparse graph codes that can be understood as a source coding analog to Gallager's [7] classical work on the effective channel capacity of bounded degree codes. We illustrated this approach by lower bounding the effective rate-distortion function of the check-regular ensemble of LDGM codes. It remains to apply our approach to other sparse code ensembles, such as LDGM families with prescribed bit and check degree distributions. In addition, our current results only guarantee that a randomly generated code will with high probability have distortion above our bound, but do not provide this guarantee for every code.

Acknowledgments

REFERENCES

[1] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley Interscience, New York, 2000.
[2] S. Ciliberti and M. Mézard. The theoretical capacity of the parity source coder. Technical report, August 2005. arXiv:cond-mat/0506652.
[3] S. Ciliberti, M. Mézard, and R. Zecchina. Message-passing algorithms for non-linear nodes and data compression. Technical report, November 2005. arXiv:cond-mat/0508723.
[4] S. Cocco, O. Dubois, J. Mandler, and R. Monasson. Rigorous decimation-based construction of ground pure states for spin-glass models on random lattices. *Phys. Rev. Letters*, 90(4), Jan. 2003.
[5] N. Creignou, H. Daudé, and O. Dubois. Approximating the satisfiability threshold of random XOR formulas. *Combinatorics, Probability and Computing*, 12:113–126, 2003.
[6] O. Dubois and J. Mandler. The 3-XORSAT threshold. In *Proc. 43rd Symp. FOCS*, pages 769–778, 2002.
[7] R. G. Gallager. *Low-density parity check codes*. MIT Press, Cambridge, MA, 1963.
[8] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford Science Publications, Clarendon Press, Oxford, 1992.
[9] M. W. Marcellin and T. R. Fischer. Trellis coded quantization of memoryless and Gauss-Markov sources. *IEEE Trans. Communications*, 38(1):82–93, 1990.
[10] E. Martinian and M. J. Wainwright. Analysis of LDGM and compound codes for lossy compression and binning. In *Workshop on Information Theory and Applications (ITA)*, February 2006. Available at arxiv:cs.IT/0602046.
[11] E. Martinian and M. J. Wainwright. Low density codes achieve the rate-distortion bound. In *Data Compression Conference*, volume 1, March 2006. Available at arxiv:cs.IT/061123.
[12] E. Martinian and M. J. Wainwright. Low density codes can achieve the Wyner-Ziv and Gelfand-Pinsker bounds. In *International Symposium on Information Theory*, July 2006. Available at arxiv:cs.IT/0605091.
[13] E. Martinian and J.S. Yedidia. Iterative quantization using codes on graphs. In *Allerton Conference on Control, Computing, and Communication*, October 2003.
[14] Y. Matsunaga and H. Yamamoto. A coding theorem for lossy data compression by LDPC codes. *IEEE Trans. Info. Theory*, 49:2225–2229, 2003.
[15] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina. Alternative solutions to diluted p-spin models and XORSAT problems. *Jour. of Statistical Physics*, 111:105, 2002.
[16] T. Murayama. Thouless-Anderson-Palmer approach for lossy compression. *Physical Review E*, 69:035105(1)–035105(4), 2004.
[17] A. J. Viterbi and J. K. Omura. Trellis encoding of memoryless discrete-time sources with a fidelity criterion. *IEEE Trans. Info. Theory*, IT-20(3):325–332, 1974.
[18] M. J. Wainwright and E. Maneva. Lossy source coding by message-passing and decimation over generalized codewords of LDGM codes. In *International Symposium on Information Theory*, Adelaide, Australia, September 2005. Available at arxiv:cs.IT/0508068.