Local Recovery in Data Compression for General Sources

Arya Mazumdar Venkat Chandar Gregory W. Wornell

***Abstract*—Source coding is concerned with optimally compressing data, so that it can be reconstructed up to a specified distortion from its compressed representation. Usually, in fixed-length compression, a sequence of n symbols (from some alphabet) is encoded to a sequence of k symbols (bits). The decoder produces an estimate of the original sequence of n symbols from the encoded bits. The *rate-distortion function* characterizes the optimal possible rate of compression allowing a given distortion in reconstruction as n grows. This function depends on the source probability distribution.**

In a *locally recoverable* decoding, to reconstruct a single symbol, only a few compressed bits are accessed. In this paper we find the limits of local recovery for rates near the rate-distortion function. For a wide set of source distributions, we show that, it is possible to compress within ϵ of the rate-distortion function such the local recoverability grows as $\Omega(\log(\frac{1}{\epsilon}))$; that is, in order to recover one source symbol, at least $\Omega(\log(\frac{1}{\epsilon}))$ bits of the compressed symbols are queried. We also show order optimal impossibility results. Similar results are provided for lossless source coding as well.

I. Introduction

Motivated by distributed storage applications, in [5], [6], the problem of designing capacity-approaching error-correcting codes with good update-efficiency and local recovery properties was introduced. The authors also made some observations regarding the analogous problem of source coding. For source codes, while no non-trivial results are presented for the update-efficiency problem, the authors point out that existing results on low density generator matrix (LDGM) codes provide a tight (up to constant factors) characterization of the trade-off between the excess rate and local recoverability for the special case of quantizing a binary symmetric source under Hamming distortion [6, Sec. VIII]. The purpose of this paper is to illustrate that the tradeoff between excess rate and local recoverability holds for more general rate-distortion problems.

Let us formalize the problem. We consider the standard rate-distortion setting. One is given a source sequence of n i.i.d. random variables $X_1, \dots, X_n$ with distribution P_X, and a (finite) distortion measure $\mathrm{d}(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the source and reconstruction alphabets, respectively. We define the distortion between a length n source sequence $\boldsymbol{x} \in \mathcal{X}^n$ and a length n reconstructed sequence $\boldsymbol{y} \in \mathcal{Y}^n$ as $\mathrm{d}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n \mathrm{d}(x_i, y_i)$. The rate-distortion function $R(D)$ denotes the minimum rate required to represent the source such that the average per-symbol distortion between the input and the reconstruction is at most D. In general, for a rate $R \geq R(D)$, an encoder maps a sequence $x_1 \dots x_n$ into a sequence $\boldsymbol{c}$ of Rn bits. The decoder maps the sequence $\boldsymbol{c}$ to a reconstructed sequence $y_1 \dots y_n$. The decoder is said to have *local recoverability* r or is r-local if each y_i is a function of at most r coordinates of $\boldsymbol{c}$.

Because of this additional constraint on the decoder, we expect that the optimal compression rate may increase beyond $R(D)$. We show that for a broad class of sources P_X and distortions $\mathrm{d}(x,y)$ (satisfying a technical condition we make precise below), a straightforward generalization of the LDGM ensemble achieves rate $R(D) + \epsilon$ and supports local recoverability $O(\log(\frac{1}{\epsilon}))$. That is, the scaling law of [6] extends beyond the binary symmetric source and Hamming distortion. Note that, the constant hidden by the big-O notation may depend on both the source distribution and the tolerable distortion. Specializing the result to the case of zero distortion, i.e., lossless source coding, we show that combining the above LDGM ensemble with a well-known sparse bit-vector data structure [1], it is possible to achieve lossless compression with rate $H(P_X) + \epsilon$ while supporting $O(\log(\frac{1}{\epsilon}))$-local recovery, where $H(P_X)$ denotes the source entropy rate (again, this result holds under a technical constraint on P_X that we make precise below). Here we consider fixed length source codes, so in general there is some probability of error, i.e., some probability that the source cannot be reconstructed correctly from the fixed length compression. With our construction, this error probability decays exponentially with the blocklength n.

Concepts related to local recovery in data compression appear in the following relevant papers. In [8], n-length sequences are stored via arithmetic coding losslessly, in a data structure that can compress any i.i.d. source with local recoverability $t \log n$ and a redundancy proportional to $\left(\frac{t}{\log n}\right)^t$ on top of the entropy. However, the data-structure of [8] also supports other database operations and the objective is to minimize query time, as opposed to the number of queried bits. On the other hand concatenating many small codes it is easy to achieve the same $t \log n$ local recoverability and redundancy equal to $\frac{\log(t \log n)}{t \log n}$ (as observed in [4]). Note that, in our results local recoverability and the redundancy can both be constants independent of n.

Section II-A proves the achievability result on lossy codes,

Arya Mazumdar is with Department of ECE, University of Minnesota, Minneapolis, MN 55455 (email: arya@umn.edu). His work was supported in part by NSF grant CCF 1318093. Venkat Chandar is with D.E. Shaw and Co., New York, NY 10036 (email: chandarvenkat@verizon.net). Gregory W. Wornell is with Massachusetts Institute of Technology. Cambridge, MA 02139. His work was supported in part by AFOSR FA9550-11-1-0183, and by NSF CCF-1319828.

and Section II-B extends the lower bound from [6] to show that the local recoverability must grow as $\Omega(\log(\frac{1}{\epsilon}))$. Section III proves the result on lossless compression. Finally, Section III-B shows that for many source models, a "shaping" penalty implies that for the lossless compression problem, codes achieving rate $H(P_X)+\epsilon$ must have $\Omega(\log(\frac{1}{\epsilon}))$-local recovery.

II. Lossy source coding

A. Rate-distortion achievability result

We start by showing that there exists a lossy code of rate $R(D) + \epsilon$ and local recoverability $O(\log(\frac{1}{\epsilon}))$. The source encoder maps a sequence $X_1 \dots X_n$ into a sequence $\boldsymbol{c}$ of Rn bits. Our goal is to show that the decoder mapping $\boldsymbol{c}$ to a reconstruction sequence $Y_1 \dots Y_n$ can be local, i.e., each Y_i is a function of only a small number of positions of $\boldsymbol{c}$.

Let P_Y denote the distribution for the reconstruction alphabet corresponding to the optimal conditional distribution of reconstruction alphabet given the source, i.e., $P_{Y|X}$. Our approach to code construction relies on the reverse channel (distribution) $P_{X|Y}$. To achieve the goal of local recovery, we use the generator matrix of an LDGM code to map from $\boldsymbol{c}$ to $Y_1 \dots Y_n$. For a good encoding, the compressed sequence will be nearly uniformly distributed, and upon multiplication by an LDGM matrix the output remains uniform. However in general, P_Y may not be uniform. Therefore, we consider an augmented reverse channel that incorporates a shaper to map a uniform distribution over ℓ bits to a distribution close to P_Y. As a result, we consider the equivalent channel from $\{0,1\}^\ell$ to $\mathcal{X}$ formed by composing the shaping map and the reverse channel $P_{X|Y}$. Our strategy is to construct an LDGM code over $\mathbb{F}_{2^\ell}$ that achieves low probability of error on this augmented reverse channel. Then we argue, via Azuma's inequality, that a good error correcting code is also a good rate-distortion code, allowing us to prove the following theorem.

Theorem 1: Assume, for a given distortion level D, that P_Y is dyadic, i.e., there exists an integer ℓ such that $2^\ell P_Y(y)$ is an integer for all $y \in \mathcal{Y}$. Then, there exists a sequence of codes, indexed by the blocklength n, achieving per-symbol distortion $D + \epsilon$, rate $R(D) + \epsilon$, and local recoverability $O(\log(\frac{1}{\epsilon}))$.

Note that the constant hidden by the big-O notation may depend on P_X and D, but does not depend on n. The reason we require P_Y to be dyadic is to avoid a "shaping" penalty, as discussed further in Section III-B.

1) Proof of Thm. 1 (sketch): Let $R(D) = I(X;Y)$ denote the mutual information (in bits) induced by the unaugmented reverse channel, and let $K = \frac{I(X;Y)(1-\epsilon)}{\ell}N$ denote the length of the compressed sequence (in $\mathbb{F}_{2^\ell}$) where N is the number of reconstruction symbols.

Consider the ensemble of LDGM codes given by a random binary $N \times K$ matrix such that each row of the matrix has nonzero values in Δ locations, chosen uniformly at random with replacement. Corresponding to each row of the matrix we compute a function chosen at random from a 2-universal family mapping $\mathbb{F}_{2^\ell}^{\Delta} \to \mathbb{F}_{2^\ell}$ (in the rare case of duplicates, rather than Δ we use a 2-universal family with the appropriate row weight). One example of such a mapping is $x_1, \dots, x_\Delta \to a_0 + \sum a_i x_i$, where $a_0, \dots, a_\Delta$ are chosen uniformly at random from $\mathbb{F}_{2^\ell}$. We denote the encoding of the LDGM code by $E(\cdot)$. The following lemma is crucial for the proof of theorem 1.

Lemma 2: The ensemble average symbol-error rate of the above LDGM code ensemble is $e^{-\alpha\Delta}$, i.e., decays exponentially with Δ, when used over the augmented reverse channel. The proof of this lemma is deferred to the appendix.

To complete the proof of Theorem 1, we apply Azuma's inequality. First, we expurgate the previously constructed LDGM code by only keeping a subset of the codewords whose pre-images are messages that are at least $Ke^{-\alpha\Delta}$ apart. If we choose $\Delta = \Omega(\log\frac{1}{\epsilon})$ this can be accomplished at a rate loss of $O(\epsilon^2)$ (one can choose a length K code achieving the Gilbert-Varshamov bound). The resulting expurgated LDGM code has average block error probability $o(1)$. With a final layer of expurgation, we construct a code with maximum (as opposed to average) error probability $o(1)$.

Since the maximum error probability is small, each codeword has a decoding region inside the typical set that is of size $2^{NH(X|Y)}$, and distinct codewords have disjoint regions. Therefore, the probability that the distortion is at most D is lower bounded by $2^{N(H(X|Y)+R\ell-O(\epsilon^2)-H(X))-o(N)} = 2^{-O(N\epsilon^2)}$. On the other hand, if the expected distortion is more than $D+\gamma\epsilon$, for a large-enough cosntant γ, by Azuma's inequality, probability that the distortion is at most D is upper-bounded by $2^{-N\gamma^2\epsilon^2/2}$. Hence, we conclude that the expected distortion is at most $D + O(\epsilon)$.

Remark 1: Using the expander graph construction in Section III, we can eliminate the ϵ from the distortion, i.e., Thm. 1 holds even if we ask for a sequence of codes achieving distortion D, rate $R(D) + \epsilon$, and $O(\log(\frac{1}{\epsilon}))$-local recovery.

B. Rate-distortion lower bound

In this section, we observe that for any data compression code of rate $R(D)$ that achieves average per-symbol distortion $D+\epsilon$, the local recoverability must be $\Omega(\log(\frac{1}{\epsilon}))$. The special case of this fundamental limit for the binary symmetric source and Hamming distortion was proved in [2], [6]. Indeed, the technique of [2], [6] of can be extended in a straightforward manner to more general rate-distortion problems. In particular, the $\Omega(\log(\frac{1}{\epsilon}))$ lower bound holds via essentially the same proof as in [2] whenever the rate-distortion function $R(D)$ is differentiable at the target distortion level D.

Theorem 3: Let $\mathcal{C}$ be a rate-distortion code with rate $R = R(D)$ achieving average distortion $D + \epsilon$, and assume that $R(D + x)$ is differentiable at $x = 0$. Then, for all sufficiently small ϵ, the average local recoverability of $\mathcal{C}$, i.e., the average (over indices) number of compressed bits that must be queried to recover a symbol in a codeword, grows as $\Omega(\log(\frac{1}{\epsilon}))$.

Proof (see also, [2]): Consider a code $\mathcal{C}$ with local recoverability r and rate $R = R(D)$, achieving average distortion $D + \epsilon$. We count the number of pairs (x^N, y^N) of source strings and codewords such that x^N is typical and $\mathrm{d}(x^N, y^N) \leq N(D + \epsilon + \delta)$ in two different ways. First, from the perspective of codewords, the number of such pairs is simply $2^{RN}\,\mathrm{Vol}(D+\epsilon+\delta)$, where $\mathrm{Vol}(\alpha)$ denotes the number of typical x^N within distortion at most αN to a given codeword. Note that, $\frac{1}{N}\log\mathrm{Vol}(D+x) \to H(X) - R(D+x)$.

Now, we count the same pairs (x^N, y^N) from the perspective of source strings. For a source string x^N, note that Azuma's inequality implies that the per-symbol distortion is bounded by $D+\epsilon+o(1)$ with high probability, i.e., almost every typical x^N has a codeword y^N within distortion $D+\epsilon+o(1)$. Because of the average local recoverability and Markov's inequality, at least $\frac{RN}{2}$ bits of the lossy compression are involved in fewer than $\frac{2r}{R}$ outputs. Therefore, the number of pairs (x^N, y^N) is lower bounded by $2^{NH(X)+\frac{RN}{2}h_{\mathrm{B}}(\frac{\delta}{rD^*})}$, where D^* denotes the maximum distortion between any pair of symbols (x,y) and $h_{\mathrm{B}}(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. Comparing these counts we have, $R(D) - R(D+\epsilon+\delta) \geq \frac{R}{2}h_{\mathrm{B}}(\frac{\delta}{rD^*})$.

To complete the proof, note that $R(D+x)$ is differentiable at $x=0$. Hence, we conclude that $(\epsilon+\delta)(-\frac{d}{dx}R(D+x)) \geq \frac{R}{2}h(\frac{\delta}{rD^*})$. Therefore, $\epsilon \geq \frac{R}{2}\frac{h(\frac{\delta}{rD^*})}{-\frac{d}{dx}R(D+x)} - \delta$. Optimizing over δ (identical to [2]), we obtain $\epsilon \geq e^{-\alpha r}$ for a suitable constant α that does not depend on N, r or ϵ. ■

In the next section we extend our results to lossless source coding, i.e., the $D = 0$ case.

III. Lossless source coding

A. Achievability result

We note that, a modification of the previous achievability scheme provides non-trivial bounds on local recovery for lossless source coding, i.e., the zero distortion case. The main result of this section is the following.

Theorem 4: Consider an i.i.d. source $X_1, \ldots, X_N$ with distribution P_X. There exists a source code that compresses $X_1, \ldots, X_N$ at rate $H(P_X) + \epsilon$ and has local recoverability $O(\log(\frac{1}{\epsilon}))$, when P_X is dyadic, and recoverability $O((\frac{1}{\epsilon}\log\frac{1}{\epsilon})$, for general P_X. The probability of error, i.e., the probability that the source cannot be encoded in a manner allowing perfect reconstruction, is at most $e^{-\alpha N}$ for a suitable constant $\alpha > 0$ depending on P_X and ϵ.
Recall, in a dyadic distribution P_X over $\mathcal{X}$ there exists an integer ℓ such that $2^{\ell}P_X(x)$ is an integer for all $x \in \mathcal{X}$. Also, the constant hidden inside the big-O notation may depend on P_X, but does not depend on the blocklength N or ϵ.

The proof of the above results is a combination of the rate-distortion construction above with a construction from [1] for storing sparse bit vectors. Specifically, we construct a local rate-distortion code for P_X that achieves per-symbol average distortion δ (under Hamming distortion) using the LDGM construction. Here, δ is a parameter we set later. Next, we store the error vector, the difference between source and reconstruction (that has on average δN nonzero values) using the expander graph data structure of [1]. Although it is not stated explicitly, the construction of [1] is capable of encoding a sparse bit vector such that every bit is recovered perfectly. Indeed, in the data-structure instead of querying a single bit, taking the majority of a bit's neighbors (in the expander graph) is guaranteed to recover the value of the source bit.

It is well-known that bipartite expanders exist with N vertices on the left, $O(K\log(\frac{N}{K}))$ vertices on the right, and degree $\Delta = O(\log(\frac{N}{K}))$ such that every subset S of the left vertices of size at most K has at least $\frac{3D}{4}|S|$ neighbors on the right [3]. Using such expanders in the construction of [1], we obtain a code capable of storing any error vector with weight at most δN with rate $O(\delta\log(\frac{1}{\delta}))$ and local recoverability $O(\log(\frac{1}{\delta}))$. Finally, any symbol of the source can be reconstructed from the rate-distortion code and the compressed error. The total compression rate is the sum of the rates of the two codes, and the local recoverability is the sum of the local recoverability of the two codes.

Proof of Thm. 4: For lossless coding, we may assume that the reconstruction alphabet $\mathcal{Y} = \mathcal{X}$, and define the distortion metric as Hamming distortion. Then, assuming that P_X is dyadic, applying Thm. 1 to the $D = 0$ case, we obtain an $O(\log(\frac{1}{\delta}))$- local code with rate $H(P_X) + \delta^2$, distortion δ^2.

We label the elements of $\mathcal{X}$ as $0, 1, \ldots, |\mathcal{X}|-1$, and view the elements of $\mathcal{X}$ as elements of $\mathbb{Z}_{|\mathcal{X}|}$, the integers mod $|\mathcal{X}|$. It now makes sense to talk about addition of elements of $\mathcal{X}$. The average Hamming distance between the reconstruction produced by the LDGM rate-distortion code and the original string $X_1, \ldots, X_N$ is $N\delta^2$. Hence from Azuma's inequality, for a suitable $\alpha > 0$, the error vector contains at most $2N\delta^2$ non-zero entries with probability $1 - e^{-\alpha N}$.

To complete the construction, we need a local code that can represent such sparse error patterns. As observed in [1] for the binary alphabet case, expander graphs provide a simple method of storing sparse error vectors with very strong locality properties. In fact, [1] proves a result stronger than what we require, as they consider querying only one bit. In our case, we use an expander graph with the following parameters. There are N vertices on the left, $O(N\delta^2\log(\frac{1}{\delta}))$ vertices on the right, and every vertex on the left has degree $d = O(\log(\frac{1}{\delta}))$. The graph is a strong expander with the property that every subset S of fewer than $4\delta^2 N$ left vertices has at least $\frac{3d|S|}{4}$ neighbors on the right. For such an expander, the encoding algorithm from [1] produces an assignment of bits (0 or 1) to the vertices on the right, with the following guarantee: for every sparse error pattern with at most $2\delta^2 N$ ones, every vertex on the left is equal to the majority among the values of its neighbors on the right. For the nonbinary case, we can extend the above expander graph construction by encoding the error pattern bit by bit, i.e., create $\log_2(|\mathcal{X}|)$ expander graphs and encode from least significant to most significant bit in separate graphs. In summary, we have a code with rate $O(\delta^2\log(\frac{1}{\delta}))$ capable of storing any error pattern with fewer than $2\delta^2 N$ errors.

We combine the above code with the code produced by Theorem 1 in the obvious manner. First, encode the original source $X_1, \ldots, X_N$ with the rate-distortion code. Using the additive structure on the source alphabet $\mathcal{X}$, we end up with a residual error pattern that, with exponentially high probability, contains at most $2\delta^2 N$ errors. We encode this error pattern using $\log_2(|\mathcal{X}|)$ expander graphs, at the cost of an additional

$O(\delta^2 \log(\frac{1}{\delta}))$ rate. To locally decode a symbol from this compression, first decode using the rate-distortion code. Then, query each of the expander graphs to determine the error from least to most significant bit, and add the error and the value of the rate-distortion reconstruction to obtain the final decoded value. This algorithm requires querying a total of $O(\log(\frac{1}{\delta})) + \log_2(|\mathcal{X}|)O(\log(\frac{1}{\delta})) = O(\log(\frac{1}{\delta}))$ bits of the compression. The compression rate is $H(P_X) + O(\delta^2 \log(\frac{1}{\delta}))$, and and the probability of error is at most $e^{-\alpha N}$. By choosing $\epsilon = O(\delta^2 \log(\frac{1}{\delta}))$ we obtain a code with rate $H(P_X) + \epsilon$ and local recoverability $O(\log \frac{1}{\epsilon})$.

Finally, if P_X is not dyadic, then we cannot use Thm. 1. Instead, we can use a random code of length $O(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$, which trivially achieves compression rate $H(P_X) + \epsilon$ (see, [9]). As described in [6], such a code can be repeated to form a length N code. Combined with the expander graph construction, this produces a code with rate $H(P_X) + O(\epsilon)$, local recoverability $O(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$, and exponentially decaying probability of error. ■

B. Lower bound

In the rest of this section we show that when the source probabilities are not all dyadic fractions, a lower bound on local recoverability for lossless coding can be obtained via a shaping argument. Note, for any function $f : \{0,1\}^\ell \to \mathcal{X}$, $\Pr\left[f(b_1, \ldots, b_\ell) = x\right] = \frac{m}{2^\ell}$ when the input bits $b_1, \ldots, b_\ell$ are uniformly distributed, i.e., the probability that f takes on any particular value is an integral multiple of $2^{-\ell}$. If P_X does not take on such values, there is a shaping penalty in using codes with local recoverability r. Let $x_0 \in \mathcal{X}$ be such, that $P_X(x_0)$ is not a multiple of $2^{-\ell}$ for any integer ℓ. Consider an arbitrary code with local recoverability r capable of storing the source $X_1, \ldots, X_n$, specified by reconstruction functions $f_i : \{0,1\}^r \to \mathcal{X}$ for each source bit. We partition the f_i into two classes, namely, S_1: those f_i such that $\Pr[f_i(U_r) = x_0] < P_X(x_0)$ and S_2 : those f_i such that $\Pr[f_i(U_r) = x_0] > P_X(x_0)$ (by assumption, equality is impossible). Here, U_r is just a placeholder for the r-bit inputs to the function f_i. Let $\delta(r) \equiv \min_{m \in \mathbb{Z}} |P_X(x_0) - m2^{-r}|$. Let the type of x_0 on S_1 denote the relative frequency of x_0 in the set of indices $i : f_i \in S_1$, in the reconstruction sequence, with a similar notion for S_2. If the Rn compressed bits are set uniformly, then using standard concentration inequalities, $\Pr\left[\text{Type of } x_0 \text{ on } S_1 \geq P_X(x_0) - \frac{\delta}{2}\right] \leq e^{-\frac{\delta^2 |S_1|}{4}}$ and $\Pr\left[\text{Type of } x_0 \text{ on } S_2 \leq P_X(x_0) + \frac{\delta}{2}\right] \leq e^{-\frac{\delta^2 |S_2|}{4}}$. If we define a "typical set" for the source as those strings whose type for x_0 is $> P_X(x_0) - \frac{\delta}{2}$ on S_1 and $< P_X(x_0) + \frac{\delta}{2}$ on S_2, then with overwhelming probability the source sample lies within this typical set. However, since $|S_1| + |S_2| = n$, at most $2^{Rn - O(\delta^2 n)}$ reconstruction strings lie within this typical set. Therefore, for high probability reconstruction to be possible, $R - O(\delta(r)^2) \geq H(P_X)$. Note that $\delta(r)$ is related to the binary expansion of $P_X(x_0)$. Specifically, the approximation error is governed up to a constant factor by the location of the first 1 in the binary expansion of $P_X(x_0)$ after the r^{th} position. For dyadic fractions, all the digits in the binary expansion are eventually 0, but for Lebesgue a.e. number in $[0,1]$, $\delta(r) = \Omega(\frac{1}{2^{2r}})$ for all sufficiently large r. That is, for almost every number in $[0,1]$, not all the digits between positions r and $2r$ in the binary expansion are 0. Substituting $\delta(r) = \Omega(\frac{1}{2^{2r}})$ we have $R \geq H(P_X) + 2^{-cr}$, for some constant c. Hence we have the following converse.

Theorem 5: For Lebesgue a.e. distribution P_X, local recoverability must grow as $\Omega(\log(\frac{1}{\epsilon}))$ to achieve a compression rate $H(P_X) + \epsilon$.

IV. Conclusion

We show, for both lossy and lossless data compression, that it is possible to compress a generic source very close to the Shannon limit (rate-distortion function and entropy respectively), while achieving a local recoverability that scales as log of one over the gap. It would be interesting to prove non-trivial lower bounds on the required local recoverability for compressing sources where the shaping argument such as in Sec. III-B does not apply, e.g., a Bernoulli(p) source where p is a dyadic fraction. Conversely, along the lines of [7], it would be interesting to show that the local recoverability does not need to scale at all with gap to the Shannon limit.

Appendix
Proof of Lemma 2

To analyze the symbol-error rate (SER), note that, by permutation symmetry of the information patterns and the code ensemble, we can consider the case when a fixed information pattern $\boldsymbol{m}$ is sent. Instead of analyzing the maximum likelihood (ML) decoder directly, we consider a combined typicality plus ML decoder (by ML decoder we mean the optimal SER decoder, as opposed to the decoder producing the most likely codeword). First, the decoding algorithm searches for all codewords of the LDGM code typical with the received string $W(E(\boldsymbol{m}))$. Here, $W(\cdot)$ denote the noisy output of the augmented reverse channel. Then, among the typical codewords, we make a decision based on ML decoding.

We split the analysis of the SER of this decoder into two cases. The first case is that an error is caused by an information pattern $\boldsymbol{m}'$ such that the Hamming distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$ is greater than $\frac{\beta K}{\Delta}$ for a parameter β to be specified later. The second case is that the Hamming distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$ is greater than $Ke^{-\alpha\Delta}$ but less than $\frac{\beta K}{\Delta}$, where α is a parameter we specify later. We will show that in both cases, the probability of error goes to zero. This lets us conclude that, with high probability, the ML decoder outputs an information pattern with Hamming distance at most $Ke^{-\alpha\Delta}$ from the sent pattern $\boldsymbol{m}$, i.e., the SER is at most $e^{-\alpha\Delta}$.

For the first case, we further assume that the channel behaves typically, that is, the input codeword and the channel output are jointly typical (we use strong typicality; specifically, we say that strings are typical / jointly typical if the deviation of each component from its mean is at most $O(N^{\frac{3}{4}})$). Note that this holds except with vanishing probability. Now, because the Hamming distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$ is so large, a simple typicality argument suffices to bound the error probability. It is extremely unlikely that $\boldsymbol{m}'$ is jointly typical with the channel output. This probability can be bounded by

$$\begin{aligned}
& \Pr(E(\boldsymbol{m}), E(\boldsymbol{m}'), W(E(\boldsymbol{m})) \text{ jointly typical}) \\
= & \sum_{\boldsymbol{s}\in T(U)} \sum_{\boldsymbol{x}\in T(\boldsymbol{s})} \sum_{\boldsymbol{s}'\in T(\boldsymbol{x})} \Pr(E(\boldsymbol{m}) = \boldsymbol{s}, E(\boldsymbol{m}') = \boldsymbol{s}') \\
& \quad \cdot \Pr(W(E(\boldsymbol{m})) = \boldsymbol{x} | E(\boldsymbol{m}) = \boldsymbol{s}) \\
\leq & \sum_{\boldsymbol{s}\in T(U)} \sum_{\boldsymbol{x}\in T(\boldsymbol{s})} |T(\boldsymbol{x})| \max_{\boldsymbol{s},\boldsymbol{s}'} \Pr(E(\boldsymbol{m}) = \boldsymbol{s}, E(\boldsymbol{m}') = \boldsymbol{s}') \\
& \quad \cdot \Pr(W(E(\boldsymbol{m})) = \boldsymbol{x} | E(\boldsymbol{m}) = \boldsymbol{s}) \\
\leq & \sum_{\boldsymbol{s}\in T(U)} \max_{\boldsymbol{x}\in T(\boldsymbol{s})} |T(\boldsymbol{x})| \max_{\boldsymbol{s},\boldsymbol{s}'} \Pr(E(\boldsymbol{m}) = \boldsymbol{s}, E(\boldsymbol{m}') = \boldsymbol{s}').
\end{aligned}$$

In the above equations, with a slight abuse of notation, $T(\boldsymbol{y})$ denotes the set of strings jointly typical with $\boldsymbol{y}$, and $T(U)$ denotes the set of strings typical for the uniform distribution. Because each output symbol is generated independently, and the function computed by each node is chosen from a 2-universal family, for our ensemble, $\max_{\boldsymbol{s},\boldsymbol{s}'} \Pr(E(\boldsymbol{m}) = \boldsymbol{s}, E(\boldsymbol{m}') = \boldsymbol{s}') = (\max_{s,s'} \Pr(E(\boldsymbol{m})(i) = s, E(\boldsymbol{m}')(i) = s'))^N \leq (2^{-\ell}(2^{-\ell} + \omega^{\Delta}))^N$, where ω denotes the fraction of positions where $\boldsymbol{m}$ and $\boldsymbol{m}'$ are identical, i.e., ω is 1 minus the relative Hamming distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$. Denote by $\mathcal{D}(\omega)$ the set of 2^{ℓ}-ary K-vectors with ω proportion of zeros. Clearly, $|\mathcal{D}(\omega)| = \binom{K}{\omega K}(2^{\ell} - 1)^{K-\omega K}$. Below, $\tilde{Y}$ denote a uniform 2^{ℓ}-ary random variable representing the LDGM codeword symbols. Applying a union bound, we see that the ensemble average probability of error between $\boldsymbol{m}$ and $\boldsymbol{m}'$ for messages $\boldsymbol{m}'$ far away from $\boldsymbol{m}$ can be upper bounded by $\Pr(E(\boldsymbol{m}) \text{ atypical}) + \Pr(E(\boldsymbol{m}), W(E(\boldsymbol{m})) \text{ jointly atypical})$ plus $\sum_{\omega} |\mathcal{D}(\omega)|(2^{-\ell}(2^{-\ell} + \omega^{\Delta}))^N |T(U)| \max_{\boldsymbol{x}} |T(\boldsymbol{x})|$,

$$\begin{aligned}
\leq & \; O(2^{-N^{\frac{1}{4}}}) + \sum_{\omega} 2^{NH(\tilde{Y}|X) + NH(\tilde{Y}) + O(N^{\frac{3}{4}})} \\
& \cdot 2^{-\ell N}(2^{-\ell} + \omega^{\Delta})^N 2^{RN(h_{\mathrm{B}}(\omega) + (1-\omega)\log(2^{\ell}-1))} \\
\leq & \; O(2^{-N^{\frac{1}{4}}}) + O(2^{N^{\frac{7}{8}}}) \max_{\omega < 1 - \frac{1}{\Delta}} 2^{NH(\tilde{Y}|X) - NH(\tilde{Y})} \\
& \cdot 2^{N\log(1 + 2^{\ell}\omega^{\Delta}) + RN(h_{\mathrm{B}}(\omega) + (1-\omega)\log(2^{\ell}-1))} = o(1),
\end{aligned}$$

as long as $\min_{\omega < 1 - \frac{1}{\Delta}} I(\tilde{Y}; X) - \frac{2^{\ell}}{\ln 2}\omega^{\Delta} - R(h_{\mathrm{B}}(\omega) + (1 - \omega)\log(2^{\ell} - 1)) > 0$.

It follows that for $R = \frac{I(X;Y)}{\ell}(1 - \epsilon)$, the above expression is positive over the range $\omega < (2^{-\ell} I(X;Y)\epsilon \ln 2)^{\frac{1}{\Delta}} = 1 - O\left(\frac{\log(\frac{1}{\epsilon})}{\Delta}\right)$, where the constant hidden in the big-O notation depends on the channel, but not on N or ϵ. Taking this one step further, when $\omega = 1 - O\left(\frac{\log(\frac{1}{\epsilon})}{\Delta}\right)$, $R(h_{\mathrm{B}}(\omega) + (1 - \omega)\log(2^{\ell} - 1)) < \frac{I(X;Y)}{2}$, so the above expression is actually positive for $\omega < 1 - \frac{\beta}{\Delta}$, for a suitable constant β depending only on the channel.

The second case is when the distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$ is relatively small. In this case, rather than proving that $E(\boldsymbol{m}')$ is unlikely to be jointly typical with the channel output, we show that under ML decoding, it is unlikely that $E(\boldsymbol{m}')$ is more probable than $E(\boldsymbol{m})$ given the observed sequence. Intuitively, because the distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$ is small, we expect that the distance between $E(\boldsymbol{m})$ and $E(\boldsymbol{m}')$ is $\Omega(\Delta)$ times larger than the Hamming distance between $\boldsymbol{m}$ and $\boldsymbol{m}'$.

Proceeding more formally, let us fix a distinguishable symbol pair $\tilde{y_1}, \tilde{y_2}$; by distinguishable, we mean that the conditional distributions $p(x|\tilde{y_1})$ and $p(x|\tilde{y_2})$ are not identical, i.e., $D(p(x|\tilde{y_1})||p(x|\tilde{y_2})) > 0$. We may assume that such a pair exists without loss of generality, because otherwise $I(X;Y) = 0$ and there is nothing to prove. Let $\boldsymbol{m}$ and $\boldsymbol{m}'$ be messages with Hamming distance δK, for some $e^{-\alpha\Delta} < \delta < \frac{1}{\Delta}$.

Let U be the number of coordinates where $E(\boldsymbol{m})$ is $\tilde{y_1}$ and $E(\boldsymbol{m})$ is $\tilde{y_2}$, i.e., $U = |\{i : E(\boldsymbol{m})_i = \tilde{y_1}, E(\boldsymbol{m}')_i = \tilde{y_2}\}|$. U is a binomial random variable since every check is generated independently in our ensemble. We have, $\mathbb{E}U \geq N(1 - (1 - \delta)^{\Delta})2^{-2\ell}$. Therefore, a standard large deviations calculation shows that $\Pr(U \leq N\frac{\delta\Delta}{3}2^{-2\ell}) \leq e^{-N2^{-2\ell}\frac{\delta\Delta}{8}}$. Assuming that $U > N\frac{\delta\Delta}{3}2^{-2\ell}$, the probability of confusing $\boldsymbol{m}$ with $\boldsymbol{m}'$ under ML decoding is at most $e^{-aN\frac{\delta\Delta}{3}2^{-2\ell}}$, where a is a positive constant (it is $C(p(x|\tilde{y_1})||p(x|\tilde{y_2}))$, the Chernoff exponent for discriminating between the distributions $p(x|\tilde{y_1})$ and $p(x|\tilde{y_2})$). Taking a union bound, we conclude that, under ML decoding, the probability of confusing $\boldsymbol{m}$ with $\boldsymbol{m}'$ is at most $2^{-\Omega(\delta\Delta N)}$, where the constant hidden in the big-O notation depends on the channel, but not on N, δ or ϵ. The number of messages $\boldsymbol{m}'$ at Hamming distance δK from $\boldsymbol{m}$ can be upper bounded by $2^{K(h_{\mathrm{B}}(\delta) + \ell\delta)} \leq 2^{O(K\delta\log(\frac{1}{\delta}))}$ for the range of δ of interest. Applying a union bound, we conclude that the probability of confusing $\boldsymbol{m}$ with any message at Hamming distance δK from $\boldsymbol{m}$ is at most $2^{-\Omega(\delta\Delta N) + O(K\delta\log(\frac{1}{\delta}))}$. Therefore, the probability of error under ML decoding goes to zero exponentially quickly with N whenever $\alpha\Delta > \log(\frac{1}{\delta})$ for a suitable constant α depending only on the channel. Taking a union bound over the fewer than K values of distance in the range $e^{-\alpha\Delta} < \delta < \frac{1}{\Delta}$, we conclude that the probability of error under ML decoding goes to zero for large N for all $\boldsymbol{m}'$ in this Hamming distance range.

REFERENCES

[1] H. Buhrman, P. B. Miltersen, J. Radhakrishnan, and S. Venkatesh. Are bitvectors optimal? *SIAM Journal on Computing*, 31(6):1723–1744, 2002.

[2] V. Chandar. *Sparse graph codes for compression, sensing, and secrecy*. PhD thesis, Massachusetts Inst. of Technology, Cambridge, MA, 2010.

[3] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the Amer. Math. Society*, 43(4):439–561, 2006.

[4] A. Makhdoumi, S.-L. Huang, Y. Polyanskiy, and M. Medard. On locally decodable source coding. *arXiv preprint arXiv:1308.5239*, 2013.

[5] A. Mazumdar, V. Chandar, and G. W. Wornell. Local recovery properties of capacity achieving codes. In *Info. Theory and Applications*, 2013.

[6] A. Mazumdar, V. Chandar, and G. W. Wornell. Update-efficiency and local repairability limits for capacity approaching codes. *Selected Areas of Communications, IEEE Journal on*, 32(5), 2014.

[7] A. Montanari and E. Mossel. Smooth compression, Gallager bound and nonlinear sparse-graph codes. In *Proc. Int. Symp. Inform. Theory*, pages 2474–2478, Toronto, Canada, July 2008.

[8] M. Patrascu. Succincter. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symp.*, pages 305–313. IEEE, 2008.

[9] Z. Zhang, E.-H. Yang, and V. K. Wei. The redundancy of source coding with a fidelity criterion I: Known statistics. *IEEE Trans. Inform. Theory*, 43(1):71–91, Jan. 1997.