The Effect of Local Decodability Constraints on Variable-Length Compression

Ashwin Pananjady and Thomas A. Courtade

Abstract—**We consider a variable-length source coding problem subject to local decodability constraints. In particular, we investigate the blocklength scaling behavior attainable by encodings of r-sparse binary sequences, under the constraint that any source bit can be correctly decoded upon probing at most d codeword bits. We consider both adaptive and non-adaptive access models, and derive upper and lower bounds that often coincide up to constant factors. Such a characterization for the fixed-blocklength analog of our problem, known as the bit probe complexity of static membership, remains unknown despite considerable attention from researchers over the last few decades. We also show that locally decodable schemes for sparse sequences are able to decode 0s (frequent source symbols) of the source with far fewer probes on average than they can decode 1s (infrequent source symbols), thus rigorizing the notion that infrequent symbols require high probe complexity, even on average. Connections to the fixed-blocklength model and to communication complexity are also briefly discussed.**

Index Terms—**Variable-length compression, local decoding, bit-probe, static membership.**

I. Introduction

EFFICIENT representation of a sequence of *source bits* by a significantly shorter sequence of *encoded bits* (i.e., a *codeword*) is the classical problem of lossless source coding, proposed by Shannon in his seminal 1948 paper [2]. It is widely known that optimal compression performance can be achieved with schemes such as Huffman codes [3] or the Lempel-Ziv universal compression algorithm [4]. However, these compression schemes suffer from the drawback that they do not support *local decodability*. Specifically, retrieving a single bit of the source sequence generally requires a decoder to access *all* of the encoded bits.

This is clearly undesirable in applications that favor retrieving selected pieces of information, rather than the entire source sequence. One such application is in bioinformatics [5], [6], in which a DNA sequence is stored as a binary string with relation to a *reference* sequence, with 1s representing single nucleotide polymorphisms (SNPs) at those positions. In SNP calling, we are interested in learning whether there is a SNP at position i. Since we are not interested in any other information about the sequence, we would ideally like to accomplish this by accessing few bits in the compressed representation of the DNA sequence. In this specific instance, local decodability is strongly motivated, since decompressing the whole genome can be prohibitively expensive from a memory standpoint.

Another example presents itself in the efficient storage of relationships among objects (e.g., relational databases [7]). Given a collection of n objects, the relationships among these objects can be represented by an undirected graph on n vertices, with the presence or absence of an edge (i, j) signifying that objects i and j are related or unrelated, respectively (e.g., friendships in social networks). One can think of representing all graphs with n vertices by sequences of $\binom{n}{2}$ bits representing all possible edges. A '1' at a given position indicates the presence of that edge, and a '0' indicates that it is absent. Thus, testing relationships between objects is accomplished by querying the value of the corresponding bit. As in SNP calling, it would be ideal to have a compressed representation of the graph which permits such queries upon accessing a small number of encoded bits.

We remark that both applications referred to above involve a source that is inherently sparse - both SNPs and the number of relationships are small compared to the total length of the sequence. Motivated in part by this, our objective in this paper is to analyze the fundamental tradeoffs between access constraints and compressibility of sparse sequences, in the context of locally decodable compression schemes. We consider a variable blocklength model, in which source sequences can be mapped to codewords of varying lengths, and that the decoder is informed of the codeword length at the start of the decoding process.

The problem of locally decodable source coding for fixed-length codes has been studied in the context of succinct data structures, in the *bit probe* [8], [9] and *cell probe* [10], [11] complexity models. In particular, locally decodable source coding in the fixed blocklength setting is an instance of the static membership problem in the bit-probe model, which can be stated as follows: Encode subsets $S \subseteq \{1, 2, \dots, n\}$ of size at most r into a data structure of *fixed* length ℓ, such that queries of the form "Is $i \in S$" for $i \in \{1, 2, \dots, n\}$ can be determined by probing (i.e., accessing) at most d bits in the data structure, either adaptively or non-adaptively. By letting S denote the set of indices where a binary sequence

Manuscript received December 8, 2015; revised September 11, 2017; accepted December 11, 2017. Date of publication February 9, 2018; date of current version March 15, 2018. This work was supported in part by the NSF Center for Science of Information under Grant CCF-0939370 and in part by NSF under Grant CCF-1528132.This paper was presented at 2015 ISIT [1].

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ashwinpm@eecs.berkeley.edu; courtade@eecs.berkeley.edu).

Communicated by E. Tuncel, Associate Editor for Source Coding.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2804441

has ones, the static membership problem considered by the bit probe model yields a fixed-blocklength, locally decodable representation of sparse sequences.

Buhrman *et al.* [12] analyzed the bit probe complexity of the static membership problem and provided the lower bound $\ell = \Omega(dr^{1-1/d}n^{1/d})$, which remains the best lower bound for a general n, r, d. They also showed, among other results, a scheme that for odd d, achieves a blocklength scaling as $\ell = O(rdn^{4/(d+1)})$. The interested reader is referred to [13] and references therein for a comprehensive survey on several improvements to these bounds (for specific regimes of r and d) that have been proposed in the literature [14]–[17]. Notably, Garg and Radhakrishnan [18] recently improved the bounds of Burhman *et al.* In particular, they provided tight bounds for the case of $d = 2$ adaptive bit probes and improved the upper bound of [12] for certain regimes of r and d: For odd d, they showed an upper bound of $\ell = \tilde{O}(dr^{1-1/d}n^{2/(d-1)})$, which still exhibits a substantial gap in the exponent compared to the lower bound $\ell = \Omega(dr^{1-1/d}n^{1/d})$. Thus, while significant progress has been made since the static membership problem was introduced in the 70s [8], tight bounds on its bit-probe complexity that hold in general have remained elusive.

Largely independent from the prior work on the bit probe model, locally decodable source coding has also received recent attention from the information theory community [19]–[22]. Closely related to the present paper is the recent work by Mazumdar *et al.* [19], which considers the design of locally decodable source codes under the bit probe model for general memoryless sources, with vanishing block-error probability. Among its other contributions are order-wise tight bounds for the case of i.i.d. Bernoulli sources, which was first considered by [21].

In this paper, we introduce and analyze the problem of variable-blocklength compression (to be defined precisely in Section II) of sparse sequences under local decodability constraints. Our model thus differs fundamentally from those appearing in both [12] and [19], [21] since such a variable-blocklength setting has not been previously studied. Considering a variable-blocklength setting is practically motivated because compressed file size is rarely fixed a-priori by the compression scheme, and file length is often recorded in metadata available to a decompressor. As we show in the sequel, our analysis of the variable blocklength case allows us to provide tight order-wise bounds on the average blocklength of the code in many cases. Also, in contrast to [21], we restrict ourselves exclusively to the *lossless* setting, in which the decoder must attain zero (and not vanishing) probability of error, which is motivated by high-fidelity applications such as SNP calling.

Our Contributions

In this paper, we introduce the problem of locally decodable source coding for sparse sequences with variable-blocklength codes. We provide non-asymptotic upper and lower bounds on the average blocklength attainable by such schemes, which are sharp up to constant factors in many cases of interest. We reiterate that the fixed-blocklength analog of this problem does not yet have such a sharp characterization.

Roughly speaking, our bounds show that the scaling of the average blocklength is given by $drn^{r/(rd+1)}$, which can be substantially lower than the best known upper bounds for the fixed blocklength regime holding for general r, d, n, which scale as $dr^{1-1/d}n^{2/d}$ [18].[1] As a corollary, we give necessary and sufficient conditions on the number of bit-probes required to achieve competitively optimal compression performance, finding that our schemes allow us to probe a small fraction of bits required by the trivial scheme that attains competitively optimal compression.

We also show upper bounds on the average probe depth of locally decodable schemes, and a lower bound on the number of probes that are required to decode 'infrequent' source symbols, on average. Our results articulate the notion that more infrequent symbols demand higher probe complexity, even on average. We also comment on connections to the fixed-blocklength model and to communication complexity.

II. Notation and Problem Setting

For an integer $k \geq 1$, we employ the shorthand notation $[k] \triangleq \{1, 2, \ldots, k\}$. We make frequent use of the conventional asymptotic notations $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot), \Theta(\cdot)$.

Throughout, we consider encodings of r-sparse binary vectors, which are simply sequences $x^n = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ having Hamming weight precisely r (we may assume without loss of generality that $r \leq n/2$). Our restriction to sequences of weight precisely r is primarily for convenience, since our arguments readily generalize to vectors having weight at most r. In some cases, we allow the sparsity parameter r to scale with n, in which case we write r_n.

The *support* of a source sequence x^n is defined to be the set of nonzero coordinates, i.e. $\mathsf{supp}(x^n) = \{i \in [n] : x_i = 1\}$. When referring to multiple distinct sequences, we use the bracket subscript notation, i.e. $x^n_{(1)}, x^n_{(2)}, \ldots$, where each $x^n_{(j)} \in \{0, 1\}^n$.

Letting $\binom{[n]}{r} \subset \{0, 1\}^n$ denote the set of r-sparse binary vectors, we assume random vectors $X^n \in \binom{[n]}{r}$ are drawn uniformly from all $\binom{n}{r}$ possibilities.

A source code (i.e., compressor) c for r-sparse vectors is an invertible mapping[2] $\mathsf{c} \colon \binom{[n]}{r} \to \{0, 1\}^*$, where $\{0, 1\}^* = \{0, 1, 00, 01, 10, \ldots\}$ denotes the set of all binary strings. Letting $\ell(b)$ denote the length of $b \in \{0, 1\}^*$, we remark that there are source codes for which the average codeword length is roughly[3]

$$\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] \approx \log \binom{n}{r} \text{ bits}, \tag{1}$$

and this is essentially best-possible, since the entropy of the source $H(X^n) = \log \binom{n}{r}$, and as shown in [23], the expected length of any one-one encoding is bounded from below by

$$\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] \geq H(X^n) - \log(H(X^n) + 1) - \log e \text{ bits}. \tag{2}$$

Indeed, the naïve scheme which lists the positions of each nonzero entry (requiring approximately $\log n$ bits each) is

[1] For a summary of our bounds and a detailed comparison of our bounds with those in the fixed blocklength setting, see Section III-B.

[2] Notice that we do not impose prefix constraints.

[3] Here and throughout, $\log(x)$ denotes the base-2 logarithm of x.

essentially optimal when $r \ll n$. However, it is not clear whether such a source code admits a decoding algorithm that, for any specified index $j \in [n]$, can recover bit x_j by probing a *bounded* number of bits in $\mathsf{c}(x^n)$. Thus, in the spirit of locally-decodable error-correcting codes [24] and the data structure counterparts in [12] and [21], we define a *variable-length* (r, d, n)*-locally decodable source code*:

Definition 1: A (r, d, n)*-locally decodable source code, or simply, an* (r, d, n) *code, consists of a mapping*

$$\mathsf{c} : \binom{[n]}{r} \to \{0, 1\}^*$$

with the property that, for each x^n*, the bit* x_j *can be recovered with knowledge of* $\ell(\mathsf{c}(x^n))$ *and by probing at most* d *bits of* $\mathsf{c}(x^n)$ *for all indices* $j \in [n]$.

In other words, we can say c is a (r, d, n)-locally decodable source code only if there exists a corresponding '(r, d, n)-local decompressor' — i.e., an algorithm that takes as input a *query index* $j \in [n]$ and the codeword length $\ell(\mathsf{c}(x^n))$, and returns the data bit x_j after accessing at most d bits of $\mathsf{c}(x^n)$. In light of this, we refer to the number d as an *access constraint* (or, decoding depth), since it bounds the number of encoded bits that the decoder probes before making a determination. In contrast to the fixed-blocklength settings that have been considered previously (cf. [12], [21], [24]), Definition 1 does not preclude variable-length encoding schemes. As mentioned above, this is motivated by practice, where data structures are usually of variable length and any access protocol is cognizant of the encoded data's length so that segmentation faults are avoided. Indeed, in computer file systems, a file is typically accessed after first reading metadata that describes the location and length of the file.

Note that our definition of an (r, d, n)-local decompressor does not distinguish between adaptive or non-adaptive bit probes. That is, a decompressor can probe entries of $\mathsf{c}(x^n)$ in an adaptive manner (where codeword locations are accessed sequentially, and the positions accessed can depend on the bit values observed during previous probes), or in a non-adaptive manner (where codeword locations accessed are determined only by the query index $j \in [n]$ and the codeword length $\ell(\mathsf{c}(x^n))$). When such a distinction is necessary, we explicitly refer to adaptive and non-adaptive (r, d, n) codes.

III. Main Results

A. Bounds on Expected Blocklength

In this section, we present lower and upper bounds on the expected blocklength achievable by variable-length source codes obeying a local decodability constraint, and give sufficient conditions for them to coincide. Proofs can be found in Section IV.

Theorem 1: The expected codeword length of any (r, d, n)*-locally decodable code with adaptive bit-probes satisfies*

$$\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] + 1 \geq \left(\frac{rd+1}{4e}\right)\left(\binom{n}{r}^{1/(rd+1)} - 1\right). \quad (3)$$

As a sanity check, we can evaluate the lower bound (3) when there is no locality constraint. In this case, a good compression scheme will be able to achieve an average codeword length of $H(X^n) = \log \binom{n}{r}$ bits, and so it suffices to take $d = \log \binom{n}{r}$ to specialize our result to the compression problem without a locality constraint.

Proposition 1: In case $d = \log \binom{n}{r}$*, the lower bound* (3) *reduces to*

$$\mathbb{E}[\ell(\mathsf{c}(X^n))] + 1 \geq \frac{R(r, n)}{4\ e} \log \binom{n}{r},$$

where $R(r, n)$ *is bounded above and below by universal constants* $\sqrt{2} - 1 \leq R(r, n) \leq 2$.

Hence, we recover the information-theoretic lower bound without locality constraints (1) (up to constant factors) in the absence of a local decodability constraint. On this note, an important consequence of Theorem 1 is that it dictates how quickly d must scale with respect to n, r in order to accommodate encoding schemes that are near-optimal in the traditionally information-theoretic sense. In the next section we quantify this tension more precisely, and establish how large d must be in order to ensure competitive optimality. Before doing this, we give general achievability results and discuss the tightness of (3).

We first present a scheme for the special case of $r = d = 1$. In addition to being a purely deterministic scheme, it also serves as an illustrative example of the random coding argument used in the proof of Theorem 2 to follow.

1) Deterministic Scheme for $r = 1$*,* $d = 1$*:*

a) Codebook construction: Define a sequence of sets $S_1, S_2, \ldots, S_{\lceil\sqrt{2n}\rceil}$, starting with $S_1 = \{1\}$, and inductively defining S_k to be the segment $\{m_{k-1} + 1, m_{k-1} + 2, \ldots, m_{k-1} + k\}$, with m_{k-1} denoting the maximum element of S_{k-1}. In other words, we have $S_1 = \{1\}$, $S_2 = \{2, 3\}$, $S_3 = \{4, 5, 6\}$, etc. We note that $S_1, S_2, \ldots, S_{\lceil\sqrt{2n}\rceil}$ are disjoint by construction and cover $[n]$ since

$$|S_1 \cup S_2 \cup \cdots \cup S_{\lceil\sqrt{2n}\rceil}| = \sum_{k=1}^{\lceil\sqrt{2n}\rceil} k \geq n + \sqrt{\frac{n}{2}}.$$

b) Encoding procedure: For $x^n \in \binom{[n]}{1}$, choose k such that $\mathsf{supp}(x^n) \subseteq S_k = \{m_k + 1, \ldots, m_k + k\}$ and encode x^n to $\mathsf{c}(x^n) = (x_{m_k+1}, x_{m_k+2}, \ldots, x_{m_k+k}) \in \{0, 1\}^k$.

Clearly, this encoding is $(1, 1, n)$-locally decodable. Indeed, upon observing the codeword length $k = \ell(\mathsf{c}(x^n))$, any query for index $j \notin S_k$ returns 0, whereas any query for index $j \in S_k$ is easily handled by probing and returning the $(j - m_{k-1})^{\text{th}}$ coordinate of $\mathsf{c}(x^n)$.

c) Performance analysis: Since $\Pr\{\mathsf{supp}(X^n) \subseteq S_k\} \leq \frac{k}{n}$, the expected codeword length for this encoding is given by

$$\begin{aligned}\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] &\leq \sum_{k=1}^{\lceil\sqrt{2n}\rceil} k\frac{k}{n} \\ &= \frac{\sqrt{2} + 6\sqrt{n} + 4\sqrt{2}n}{6\sqrt{n}} + O(1) \\ &= \frac{2\sqrt{2}}{3}\sqrt{n} + O(1/\sqrt{n}). \quad (4)\end{aligned}$$

Note that the expected length is within a constant of what is specified by Theorem 1. In fact, for $r = d = 1$, the lower bound in Theorem 1 can actually be improved[4] to $\mathbb{E}\left[\ell(\mathsf{C}(X^n))\right] \geq \frac{2\sqrt{2}}{3}\sqrt{n} - o(\sqrt{n})$, thereby completely characterizing the scaling behavior of an optimal $(1,1,n)$-locally decodable code as $\mathbb{E}\left[\ell(\mathsf{C}(X^n))\right] \sim \frac{2\sqrt{2}}{3}\sqrt{n}$.

Having observed that the lower bound (3) exhibits the correct scaling behavior in the $r = d = 1$ setting, we now turn to the general case.

Theorem 2: Suppose n, r *is such that the source entropy satisfies* $\log\binom{n}{r} \geq 12$. *For any* $d \geq 1$, *there exists a non-adaptive* (r,d,n)*-locally decodable code* C *with average codeword length*

$$\mathbb{E}\left[\ell(\mathsf{C}(X^n))\right] \leq 250(rd+1)\left(r^r\binom{n}{r}\right)^{1/(rd+1)}. \qquad (5)$$

A few remarks are in order. First, Theorem 1 is a converse result for adaptive schemes, while Theorem 2 is an achievability result for non-adaptive schemes. We will see in the examples that follow that these bounds often coincide (up to constant factors), showing that adaptivity provides at most constant-factor improvement in such cases.[5] Second, we note that both Theorem 1 and Theorem 2 are non-asymptotic in nature (see Remark 1 for a clarification on the constant factors involved). While Theorem 1 holds for any choice of parameters r, d, n, Theorem 2 requires that the source entropy be modestly large. Implications of the theorems together, however, become most crisp when $n \to \infty$, and r, d are functions of n. We illustrate this with some examples.

Example 1: First take $n \to \infty$ and r, d fixed (i.e., not depending on n). In this case, we find that the blocklength of an optimal sequence $\{\mathsf{C}_n^\star\}$ of (r,d,n) codes scales as $\mathbb{E}\left[\ell(\mathsf{C}_n^\star(X^n))\right] = \Theta(n^{r/(rd+1)})$. Hence, when r, d are fixed, performance scales poorly relative to the information-theoretic lower bound without locality constraints, of $\Theta(\log n)$. In fact, an *exponentially* longer codeword length is required on average! It is also worthwhile to note that the blocklength scaling behavior in the bit-probe model (discussed in Section I) remains a longstanding open problem, even in this setting of arbitrary fixed r, d and $n \to \infty$, with the best bounds differing in the exponent by a substantial margin (i.e., exponents in best upper and lower bounds differ by roughly a factor of two [18]).

Example 2: Consider the setting where $r_n = n^\epsilon$ and $d_n = \delta \log n$, where $\epsilon, \delta > 0$ are fixed constants. Then it is a straightforward calculation using (3) and (5) to see that any optimal sequence $\{\mathsf{C}_n^\star\}$ of (r_n, d_n, n)-locally decodable codes satisfies

$$C_1(2^{(1-\epsilon)/\delta} - 1) \leq \frac{\mathbb{E}\left[\ell(\mathsf{C}_n^\star(X^n))\right]}{\delta n^\epsilon \log n} \leq C_2 2^{1/\delta} \quad \text{as } n \to \infty,$$

where C_1 and C_2 are absolute constants. Thus, up to constant factors, the blocklength scaling behavior of optimal codes in this regime is $\delta n^\epsilon \log n$, and the decoder will probe a fraction of the codeword proportional to $1/r_n$ in worst case. Contrast this with the trivial encoding scheme that simply stores the position of each '1'; the natural decoder based on binary search would require roughly $\log(r_n) \cdot \log(n)$ probes in worst case.

Example 3: If we now parameterize $n_m = \binom{m}{2}$, $r_m = (1+\epsilon)\frac{\ln m}{m}\binom{m}{2}$ and $d_m = \delta \log m$, then as $m \to \infty$ any optimal sequence $\{\mathsf{C}_m^\star\}$ of (r_m, d_m, n_m)-locally decodable codes will satisfy

$$C_1(2^{1/\delta} - 1) \leq \frac{\mathbb{E}\left[\ell(\mathsf{C}_m^\star(X^{n_m}))\right]}{r_m d_m} \leq C_2 2^{2/\delta}.$$

This particular choice of parameters can be interpreted as encoding a random graph on m vertices with $(1+\epsilon)\frac{\ln m}{m}\binom{m}{2}$ edges. Since $\frac{\ln m}{m}$ is the threshold for connectivity, this graph is connected with high probability for $\epsilon > 0$. Now, querying whether two vertices are connected in this graph corresponds to querying a bit of X^{n_m}. In order to accomplish this in time that grows logarithmically in the number of vertices requires average blocklength of order $rd = \Theta\left(m\log^2(m)\right)$.

In the latter two examples, average blocklength scales $r_n d_n = \Theta(\log\binom{n}{r_n})$, which is within constant factors of the information-theoretic lower bound without locality constraints (i.e., competitively optimal). In both cases, we chose $r_n d_n = \Omega(\log\binom{n}{r_n})$ in order to achieve this scaling. Thus, it is natural to ask: *do there exist competitively optimal schemes with* $r_n d_n = o(\log\binom{n}{r_n})$*?* The answer to this question is negative, and is the focus of the next section. However, before we proceed, we unify the above examples under the following straightforward corollary of Theorems 1 and 2:

Corollary 1: If $\log\binom{n}{r_n} = \Omega(r_n d_n)$ *and* $d_n = \Omega(\log r_n)$, *then any optimal sequence of* (r_n, d_n, n)*-locally decodable codes* $\{\mathsf{C}_n^\star\}$ *satisfies*

$$\mathbb{E}\left[\ell(\mathsf{C}_n^\star(X^n))\right] = \Theta\left(r_n d_n \binom{n}{r_n}^{1/(r_n d_n + 1)}\right).$$

B. Summary and Comparison

Having seen instantiations of our results for various examples, it is now instructive to see how our results compare with those of fixed blocklength locally decodable codes. We provide such a comparison in Table I. We have deliberately simplified some of the expressions in order to make the comparison apparent. Let us parse this table row by row.

In the most general case (for arbitrary parameters (r, d, n)), our upper bound provides improvement over the best-known fixed-blocklength scheme, as seen by the simple inequality

$$\left(\frac{n}{r}\right)^{1/d} n^{1/d} > \left(\frac{n}{r}\right)^{r/(rd+1)} r^{r/(rd+1)}.$$

Now consider the case where r and d are constants (a very sparse source), and the asymptotics are with respect to n. In this regime, also known to be the difficult regime for the fixed blocklength case, we see the starkest difference between our bounds and the bounds for fixed blocklength codes. In particular, while our lower and upper bounds are sharp, the lower and upper bounds for fixed blocklength codes differ by a multiplicative factor of $n^{1/d}$. Moreover, our

[4] The proof is roughly the same as the proof of Theorem 1, but the various bounds can be improved by particularizing to $r = d = 1$.

[5] It is important to note that a non-adaptive scheme in the variable blocklength setting may still "adapt" depending on the codeword length it observes, in contrast to the fixed blocklength setting.

TABLE I

BLOCKLENGTH SCALING FOR FIXED BLOCKLENGTH AND VARIABLE BLOCKLENGTH (r, d, n) CODES. (CONSTANT AND LOGARITHMIC PREFACTORS ARE OMITTED FOR CONCISENESS)

	Fixed Blocklength		**Variable Blocklength (This paper)**	
	Lower Bound	Upper bound	Lower Bound	Upper bound
General (r, d, n)	$dr\left(\frac{n}{r}\right)^{1/d}$ [12]	$dr\left(\frac{n}{r}\right)^{1/d} n^{1/d}$ [18]	$dr\left(\left(\frac{n}{r}\right)^{r/(rd+1)} - 1\right)$	$dr\left(\frac{n}{r}\right)^{r/(rd+1)} r^{r/(rd+1)}$
(r, d) fixed	$n^{1/d}$ [12]	$n^{2/d}$ [18]	$n^{r/(rd+1)}$	$n^{r/(rd+1)}$
$r = o(\log n)$, $d = \Theta(\log(r))$	$dr\left(\frac{n}{r}\right)^{1/d}$ [12]	$dr\left(\frac{n}{r}\right)^{1/d} n^{1/d}$ [18]	$dr\left(\frac{n}{r}\right)^{r/(rd+1)}$	$dr\left(\frac{n}{r}\right)^{r/(rd+1)}$
$d = \Omega(\log\log\binom{n}{r})$	$dr\left(\frac{n}{r}\right)^{1/d}$ [12]	$dr\left(\frac{n}{r}\right)^{1/d}$ [12]	$dr\left(\frac{n}{r}\right)^{r/(rd+1)}$	$dr\left(\frac{n}{r}\right)^{r/(rd+1)}$

upper bound is asymptotically much better than the upper bound for fixed blocklength codes, as is evident from the relation

$$n^{2/d} \gg n^{r/(rd+1)}.$$

Moreover, since

$$n^{1/d} \ll n^{r/(rd+1)},$$

our upper bounds exhibit a scaling behaviour that is asymptotically smaller than the lower bound for fixed blocklength codes.

An intermediate regime is one where $r = o(\log n)$, corresponding to graph compression below the connectivity threshold (see the example above). Evaluating the bounds in this regime by choosing $d = \Theta(\log r)$, we again see that there is a multiplicative gap of $n^{1/d}$ between the lower and upper bounds in the fixed blocklength setting, while our bounds are sharp. Moreover, since

$$\left(\frac{n}{r}\right)^{1/d} n^{1/d} \gg \left(\frac{n}{r}\right)^{r/(rd+1)},$$

we obtain improvements in achievability in this regime. Also, as before, our upper bound is strictly smaller than the lower bound for fixed blocklength compression.

The bounds for fixed blocklength and variable blocklength codes are most comparable when d scales sufficiently fast relative to (n, r), i.e., $d = \Omega\left(\log\log\binom{n}{r}\right)$. In this regime, the lower and upper bounds for both fixed- and variable-blocklength codes match up to constant factors, and the upper bound for fixed-blocklength codes is achieved by a different scheme than the first three cases. However, it is worth noting that since

$$\left(\frac{n}{r}\right)^{1/d} > \left(\frac{n}{r}\right)^{r/(rd+1)},$$

we still obtain slight improvements in blocklength scaling behavior, which diminish when r grows sufficiently rapidly.

C. Local Decodability and Competitive Optimality

We now provide necessary conditions for competitive optimality. To this end, we define:

Definition 2: For a sequence of integers $\{r_n\}_{n\geq 1}$ a sequence of encoders

$$\mathsf{c}_n : \binom{[n]}{r_n} \to \{0, 1\}^* \quad n \geq 1$$

is said to be competitively optimal *if*

$$\limsup_{n\to\infty} \frac{\mathbb{E}[\ell(\mathsf{c}_n(X^n))]}{\log\binom{n}{r_n}} = O(1).$$

In other words, competitively optimal schemes attain compression rates within a constant factor of the information theoretic lower bound (without locality constraints) $\log\binom{n}{r_n}$ for large enough n.

From Theorem 1, it is possible to deduce the following necessary condition for competitive optimality:

Theorem 3: If $\{\mathsf{c}_n\}$ is a competitively optimal sequence of (r_n, d_n, n)-locally decodable codes, then $r_n d_n = \Omega\left(\log\binom{n}{r_n}\right)$.

In other words, we cannot expect to attain competitive optimality when r_n and d_n are simultaneously small relative to the source entropy (note the contrast to the sufficient conditions in Corollary 1). This relationship can be somewhat complicated since the source entropy generally depends on both n and r_n. However, when the source sequence is modestly sparse (i.e., $r_n = O(n^{1-\epsilon})$ for some $\epsilon > 0$), then the explicit dependence on r_n in Theorem 3 can be eliminated to obtain the following condition:

Corollary 2: If $r_n = O(n^{1-\epsilon})$ for some $\epsilon > 0$, then there exists a competitively optimal sequence of (r_n, d_n, n)-locally decodable codes if and only if $d_n = \Omega(\log n)$.

In contrast to Corollary 2, if $r_n = \Theta(n)$, the information theoretic lower bound without locality constraints is $\log\binom{n}{r_n} = \Theta(n)$, and the identity encoding $\mathsf{c}_n(x^n) = x^n$ is competitively optimal, with all source bits being decodable with $d_n = 1$ probes. We also remark that in the regime in which $r = \Omega(n)$, compression performance is characterized by the constant factors that the notion competitive optimality hides. However, our focus in this paper is on the sparse regime when $r = o(n)$; a discussion of the linear regime can be found in [19].

We now turn to the question of whether it is possible to design locally decodable codes which exhibit an even stronger sense of locality on average.

D. Average Probe Depth of Local Decoding

The results in the paper so far have addressed the problem of compression with a bound on the worst case probe depth over all choices of bits of the source that we wish to decode. However, we could now ask how the *average probe depth* behaves. In this section, we introduce two notions of "average" probe depth.

The first definition is a natural one – we consider the average probe depth to recover any bit of the source. More precisely, letting $\mathsf{d}(i, x^n)$ denote the number of probes used to determine the value of bit i in sequence x^n, we define:

Definition 3: The average probe depth D *to recover any bit of the source is defined as the probe depth averaged over all sequences and all possible query bits, i.e.,*

$$\begin{aligned}\mathsf{D} &\triangleq \mathbb{E}_{x^n}\mathbb{E}_{i\in[n]}\mathsf{d}(i, x^n)\\ &= \frac{1}{\binom{n}{r}}\sum_{x^n}\frac{1}{n}\sum_{i\in[n]}\mathsf{d}(i, x^n),\end{aligned}$$

Our result for this notion of average probe depth shows that there are (r, d, n)-locally decodable codes with performance suggested by Theorem 2 for which D is small.

Theorem 4: Suppose

1) $r_n d_n = o(n)$ with $\log\binom{n}{r_n} \geq 12$; and

2) $d_n \leq (1-\epsilon)\frac{\log n}{\log(Ce)}$ for some fixed $\epsilon > 0$ and $C > 1$.

Then there exists a sequence of (r_n, d_n, n) codes with expected length satisfying (5)*, for which*

$$\mathsf{D} \leq 1 + \frac{1}{C-1} + o(1)$$

bits. Here the $o(1)$ term goes to 0 *as $n \to \infty$.*

Theorem 4 essentially states that although the maximum probe depth to determine any bit in the source is d_n, provided d_n and r_n are small enough, there exists a scheme with close to optimal expected length in which we can probe only a small number of codeword bits on average to determine any bit. For instance, if d_n is a constant d and $r_n = o(n)$, then we can get away with just over 1 probe on average to determine any source bit, although we are allowed a maximum of d probes.

As it turns out, the reduction in average probe depth achieved by this scheme is a consequence of the sparsity of the source, since if a source bit is 0, we only need to probe a small number of codeword bits to determine it. A related question is the following: if we we restrict our attention to the effort required to decode infrequent symbols, is this still small on average? To answer this we require a different notion of average probe depth, which is defined below.

Definition 4: The average 1*-probe depth* D^1 *to recover* 1*s in the source is defined as the probe depth averaged over all bits that take the value* 1 *in the source sequence, i.e.,*

$$\begin{aligned}\mathsf{D}^1 &\triangleq \mathbb{E}_{x^n}\mathbb{E}_{i\in\mathsf{supp}(x^n)}\mathsf{d}(i, x^n)\\ &= \frac{1}{\binom{n}{r}}\sum_{x^n}\frac{1}{r}\sum_{i\in\mathsf{supp}(x^n)}\mathsf{d}(i, x^n).\end{aligned}$$

In order to state our results for this notion of average probe depth, we need the following precise definition of "good" locally decodable codes.

Definition 5: An α-optimal (r, d, n)-locally decodable code is one whose expected length is within a multiplicative factor $\alpha \geq 1$ of the lower bound (3)*. In other words, it is a code with expected length satisfying:*

$$\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] + 1 \leq \alpha\left(\frac{rd+1}{4e}\right)\left(\binom{n}{r}^{1/(rd+1)} - 1\right).$$

We are now ready to state our theorem, which is a converse result for 1-probe depth of α-optimal codes.

Theorem 5: Suppose $r_n d_n = o\left(\log\binom{n}{r}\right)$. Then for any α-optimal (r_n, d_n, n)-locally decodable code,

$$\mathsf{D}^1 = \Omega_\alpha(d_n), \tag{6}$$

where the prefactors in the lower bound (6) *are a function of α.*

A few remarks are now in order. Firstly, from Corollary 1 we know that provided $\log\binom{n}{r_n} = \Omega(r_n d_n)$ and $d_n = \Omega(\log r_n)$, the upper bound (5) and lower bound (3) for the expected length of (r, d, n)-locally decodable codes coincide up to a constant factor. In other words, the code of Theorem 2 is α-optimal under these conditions. Theorems 4 and 5 now clearly bring out the difference in the difficulty of decoding frequent and infrequent source symbols. In essence, the theorems convey that if a bit is 0, there are locally decodable codes for which we require a small number of probes to decode it on average, while if it is a 1, we require essentially d probes, our maximum budget, for any code that we use.

This was precisely the intuition behind the lower bound of [16] for maximum probe depth - that sparser symbols require more probes in worst case. Theorems 4 and 5 show that the same intuition is correct even in an average sense.

As mentioned in the introduction, a large portion of the literature related to this problem considers fixed-blocklength locally decodable codes. In the next section, we draw connections from our model to the fixed blocklength model by considering the general problem of compression with headers.

E. Using Headers

In the variable blocklength formulation, the length of the codeword is available as side information to the decoder; the local decodability constraint is imposed while probing the codeword itself. In principle, the side information could be viewed as a header that the decoder reads in advance, but whose contents are severely constrained to express only the codeword length.

One can now ask how things change if we were able to freely encode information in the bits of the header. The following definition makes this precise.

Definition 6: A $(r, d, n, \ell)_h$-locally decodable code is a mapping $\binom{[n]}{r} \to \{0, 1\}^\ell$ of the source sequences to codewords of fixed length ℓ, with the property that any bit of the source sequence x^n can be recovered by reading a header of h bits, and by probing d bits of $\mathsf{c}(x^n)$.

We have the following achievability result for such codes.

Theorem 6: There exists a $(r, d, n, \ell)_h$-locally decodable code with $h = \log\log\binom{n}{r} - \log d$ bits and codewords of fixed length $\ell = 3drn^{1/d}$ bits.

Conceptually, these $(r, d, n, \ell)_h$-locally decodable codes can be viewed as an intermediate construction lying between the variable blocklength codes introduced in this paper, and their fixed blocklength counterparts. When the header is allowed to encode arbitrary information, the performance of these codes coincides with the upper bound of Buhrman *et al.* [12] referenced in the fourth row of Table I: if the bits of the header are counted in our probing budget, then we obtain a fixed-blocklength code with blocklength scaling as $drn^{1/d}$ provided $d \geq \Omega(\log\log\binom{n}{r})$.

IV. Proofs of Main Results

In this section, we provide the proofs of Theorems 1 through 6. Proofs of some technical lemmas that are used below are deferred to the appendices. For convenience, we recall the following standard inequalities which are used repeatedly throughout the proofs without explicit mention:

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \frac{n^k}{k!} \leq \left(\frac{ne}{k}\right)^k.$$

Proof of Theorem 1: Our proof begins with the notion of a "decoding transcript" introduced in [12, Th. 6]. Let c be a (r, d, n)-locally decodable code. For a source sequence x^n, let $\mathsf{c}_q(x^n)$ denote the qth coordinate of the codeword $\mathsf{c}(x^n)$, and define the *transcript*

$$\begin{aligned} T_i^k \triangleq \big\{\big(q, \mathsf{c}_q(x^n_{(i)})\big) : \ell(\mathsf{c}(x^n_{(i)})) = k, &\text{ and location } q \text{ of } \mathsf{c}(x^n_{(i)}) \\ &\text{is accessed to determine } x_j \text{ for some } j \in \mathsf{supp}(x^n_{(i)})\big\}, \end{aligned} \tag{7}$$

where we have abused notation slightly by letting x_j denote the jth coordinate of sequence $x^n_{(i)}$. Note that each transcript T_i^k is a subset of $[k] \times \{0, 1\}$ of size at most rd, since $|\mathsf{supp}(x^n_{(i)})| = r$ and the decoder makes at most d probes in response to a query. Also note that for $i \neq i'$, $T_i^k \not\subseteq T_{i'}^k$. To see this, assume the contrary, that $T_i^k \subseteq T_{i'}^k$ for some $i \neq i'$. Let the encoded source word be $x^n_{(i')}$. If we now query the value of x_j for $j \in \mathsf{supp}(x^n_{(i)}) \setminus \mathsf{supp}(x^n_{(i')})$, we see that the decoder makes an error, establishing the contradiction.

Since for fixed k, the T_i^ks are not subsets of one another, an application of the LYM inequality [25] (also provided in Lemma 6, Appendix A) yields

$$\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \leq \max_{\upsilon \leq rd} \binom{2k}{\upsilon} \quad \text{for each } k. \tag{8}$$

In light of (8), the average codeword length must satisfy

$$\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] \geq \sum_{k=1}^{M(n,r,d)} k \frac{\max_{\upsilon \leq rd} \binom{2k}{\upsilon}}{\binom{n}{r}}, \tag{9}$$

where $M(n, r, d)$ is the largest integer satisfying

$$\sum_{k=1}^{M(n,r,d)+1} \max_{\upsilon \leq rd} \binom{2k}{\upsilon} > \binom{n}{r} \geq \sum_{k=1}^{M(n,r,d)} \max_{\upsilon \leq rd} \binom{2k}{\upsilon}. \tag{10}$$

Now define the probability distribution

$$Q(k) = \frac{\max_{\upsilon \leq rd} \binom{2k}{\upsilon}}{\binom{n}{r}} \quad \text{for } 1 \leq k \leq M(n, r, d)$$

and $Q(M(n, r, d) + 1) = 1 - \sum_{k=1}^{M(n,r,d)} Q(k)$. Since $Q(k) \leq Q(k + 1)$ for $k < M(n, r, d)$ by definition, we can conclude

$$\mathbb{E}\left[\ell(\mathsf{c}(X^n))\right] \geq \sum_{k=1}^{M(n,r,d)+1} k \cdot Q(k) \geq \frac{M(n, r, d) + 1}{2}. \tag{11}$$

Toward evaluating (11), we need the following technical estimate, which is proved in Appendix A.

Lemma 1: For all $M, \upsilon \geq 1$,

$$\sum_{k=1}^{M} \max_{i \leq \upsilon} \binom{2k}{i} \leq (\upsilon + 1)^{1/2} 2^{\upsilon} \frac{\left(M + 2 + \frac{\upsilon+1}{2e}\right)^{\upsilon+1}}{(\upsilon + 1)!}. \tag{12}$$

Identifying $M \leftarrow M(n, r, d) + 1$ and $\upsilon \leftarrow rd$ in (12), the first inequality in (10) can be rearranged to conclude that

$$M(n, r, d) + 3 \geq \left(\binom{n}{r}^{1/(rd+1)} - 1\right)\left(\frac{rd + 1}{2e}\right),$$

where we have made use of the lower bound $n! \geq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$. Recalling (11) proves the desired inequality. ■

Proof of Proposition 1: Let us look at the ratio

$$R(r, n) := \frac{r \log\binom{n}{r} + 1}{\log\binom{n}{r}} \left(\binom{n}{r}^{1/(r \log\binom{n}{r}+1)} - 1\right)$$

for the nontrivial regime of parameters $n \geq r + 1$ and $r \geq 1$. In case $n = r$ or $r = 0$, the lower bound evaluates to zero, rendering the claim trivially true.

Observe

$$\begin{aligned} 2\, r &\left(\binom{n}{r}^{1/(r \log\binom{n}{r}+1)} - 1\right) \\ &\geq R(r, n) \geq r \left(\binom{n}{r}^{1/(r \log\binom{n}{r}+1)} - 1\right), \end{aligned}$$

so we would like to show that the function $r\left(\binom{n}{r}^{1/(r \log\binom{n}{r}+1)} - 1\right)$ can be bounded above and below by constants.

To this end, note that the function $f : x \mapsto x^{1/(r\log(x)+1)}$ is increasing in $x > 0$. Indeed,

$$f'(x) = \frac{x^{-\frac{r \log x}{1 + r \log x}}}{(1 + r \log x)^2} \geq 0.$$

So, for $n \geq r + 1$ we have

$$\begin{aligned} (r + 1)&^{1/(r\log(r+1)+1)} \\ &\leq \binom{n}{r}^{1/(r \log\binom{n}{r}+1)} \leq \lim_{x\to\infty} x^{1/(r \log x+1)} = 2^{1/r}, \end{aligned}$$

giving the upper bound

$$R(r,n) \le 2\, r(2^{1/r} - 1) \le 2,$$

where the second inequality holds for $r \ge 1$. Likewise, the lower bound yields

$$R(r,n) \ge r\left((r+1)^{1/(r\log(r+1)+1)} - 1\right) \ge \sqrt{2} - 1.$$

■

Proof of Theorem 2: The proof is by a random coding argument, but it is important to note that standard typicality arguments are not applicable here since they do not support local decodability. Briefly, the idea behind our encoding scheme is to first encode some information about $\mathsf{supp}(\mathsf{c}(x^n))$ into the codeword length, and then carefully encode the remaining information so that bit x_j can be recovered by computing the binary AND of d encoded bits. A precise description of the codebook generation and decoding procedure is given below, along with an illustrative example in Figure 1.

A. Codebook Construction

For $k = rd+1, rd+2, \ldots$ choose a subset $S_k \subseteq [n]$ of size $\frac{1}{4}\frac{\binom{k}{d}}{\binom{rd}{d}}$ uniformly at random from all such subsets,[6] and set $S_k = [n]$ if $\frac{1}{4}\frac{\binom{k}{d}}{\binom{rd}{d}} > n$. For each $j \in S_k$, choose a subset $T_{j,k} \subseteq [k]$ of size d independently and uniformly from all such subsets. All subsets are made available to both encoder and decoder.

For a sequence $x^n \in \binom{[n]}{r}$, let $k(x^n)$ denote the smallest integer k such that the following two conditions hold:

(C1) $\mathsf{supp}(x^n) \subseteq S_k$; and
(C2) $T_{j,k} \not\subseteq \cup_{i \in \mathsf{supp}(x^n)} T_{i,k}$ for all $j \in S_k \setminus \mathsf{supp}(x^n)$.

B. Encoding Procedure

A sequence $x^n \in \binom{[n]}{r}$ is encoded to a codeword $\mathsf{c}(x^n)$ of length $k(x^n)$ satisfying

$$\mathsf{supp}(\mathsf{c}(x^n)) = \cup_{i \in \mathsf{supp}(x^n)} T_{i,k(x^n)}.$$

In other words, x^n is encoded to a vector of length $k(x^n)$, which has 1's in all positions $j \in T_{i,k(x^n)}$ if and only if $x_i = 1$.

C. Decoding Procedure

On observing the length of codeword $\mathsf{c}(x^n) = (c_1, c_2, \ldots, c_{\ell(\mathsf{c}(x^n))})$, determine bit x_j as follows:

(1) If $j \notin S_{\ell(\mathsf{c}(x^n))}$, declare $x_j = 0$; else
(2) If $j \in S_{\ell(\mathsf{c}(x^n))}$, declare $x_j = \wedge_{i \in T_{j,k(x^n)}} c_i$, where '$\wedge$' denotes binary AND.

By the nature of the codebook construction, it is clear that the decoder (i) will never make an error; and (ii) satisfies the non-adaptive d-local decodability constraint.

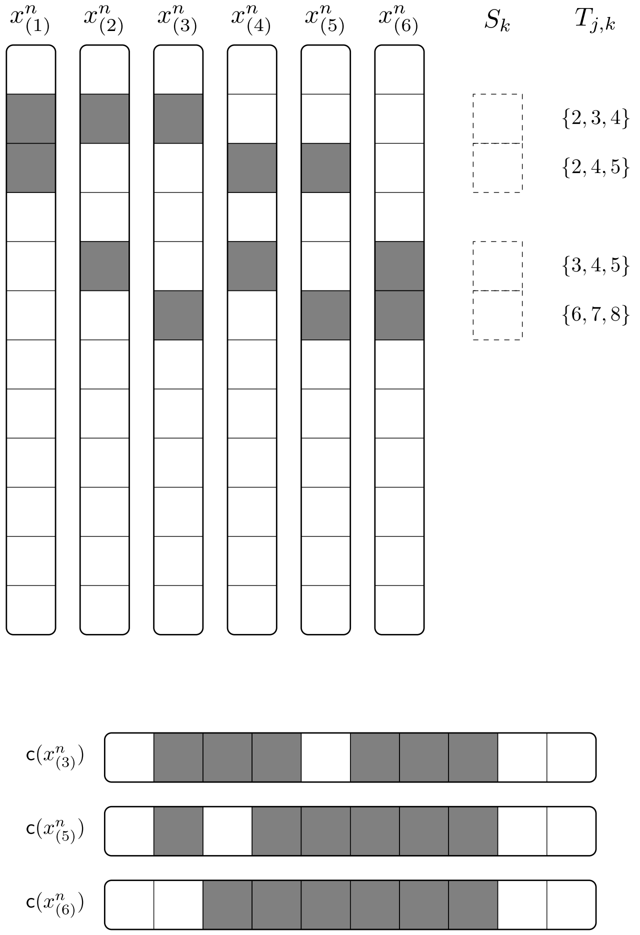

Fig. 1. Encoding of 2-sparse source sequences of length $n = 12$ using a codeword of length $k = 10$, with at most $d = 3$ bit probes. As a result, $|S_k| = \frac{r}{r+1}\binom{k}{d}/\binom{rd}{d} = 4$. The source sequences $x^n_{(1)}, \ldots, x^n_{(6)}$ represent those that satisfy condition (C1) of the encoding criterion for $S_k = \{2,3,5,6\}$. Condition (C2) is only satisfied by $x^n_{(3)}, x^n_{(5)}, x^n_{(6)}$, which are encoded as shown. If the bit to be decoded $j \in [n] \setminus S_k = \{1,4,7,8,9,10,11,12\}$, then the decoder outputs 0 without probing the bits of the codeword. If $j \in S_k$, then the decoder probes positions $T_{j,k}$ of the codeword and returns the AND of the bits (shaded blocks correspond to 1's, unshaded blocks signify 0's).

D. Performance Analysis

Recall that each codeword length k defined some region of encoding S_k, where $|S_k| = \min\left\{n, \frac{1}{4}\frac{\binom{k}{d}}{\binom{rd}{d}}\right\}$. To show a bound on the expected codeword length, fix an arbitrary sequence x^n and define the events

$$\mathcal{E}_{k,1} = \{\mathsf{supp}(x^n) \subseteq S_k\}, \quad \text{and}$$
$$\mathcal{E}_{k,2} = \{T_{j,k} \not\subseteq \cup_{i \in \mathsf{supp}(x^n)} T_{i,k} \quad \text{for all } j \in S_k \setminus \mathsf{supp}(x^n)\}.$$

By independence of the sets used in the codebook construction, we have

$$\mathbb{E}_{\mathcal{C}}\left[\ell(\mathsf{c}(x^n))\right] = \sum_{k \ge rd+1} k \Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\} \times \prod_{j=rd+1}^{k-1} \left(1 - \Pr\{\mathcal{E}_{j,1} \cap \mathcal{E}_{j,2}\}\right), \quad (13)$$

where $\mathbb{E}_{\mathcal{C}}[\cdot]$ denotes expectation over the ensemble of random codebooks. Importantly, we note that (13) is a decreasing

[6] Floor and ceiling operators are omitted for clarity of presentation.

function of $\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\}$ for each k. Therefore, in order to upper bound (13), we lower bound $\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\}$. To that end, observe that

$$\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\} = \Pr\{\mathcal{E}_{k,1}\} \Pr\{\mathcal{E}_{k,2} | \mathcal{E}_{k,1}\}$$

where the conditional probability $\Pr\{\mathcal{E}_{k,2} | \mathcal{E}_{k,1}\}$ can be bounded from below by a simple union bound:

$$\begin{aligned}\Pr\{\mathcal{E}_{k,2} | \mathcal{E}_{k,1}\} &\geq 1 - \sum_{j \in S_k \setminus \mathsf{supp}(x^n)} \Pr\{T_{j,k} \subseteq \cup_{i \in \mathsf{supp}(x^n)} T_{i,k}\} \\ &\geq 1 - |S_k| \frac{\binom{rd}{d}}{\binom{k}{d}}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\} &\geq \begin{cases} \frac{1}{\binom{n}{r}} \binom{|S_k|}{r} \left(1 - |S_k| \frac{\binom{rd}{d}}{\binom{k}{d}}\right), & \text{if } |S_k| < n \\ 1 - n \frac{\binom{rd}{d}}{\binom{k}{d}}, & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{\binom{n}{r}} \frac{3}{4} \binom{\frac{1}{4} \frac{\binom{k}{d}}{\binom{rd}{d}}}{r}, & \text{if } |S_k| < n, \\ 1 - n \frac{\binom{rd}{d}}{\binom{k}{d}}, & \text{otherwise,} \end{cases} \qquad (14)\end{aligned}$$

where (14) follows from substituting the value of $|S_k|$. The challenge of the proof is to now carefully bound (14) and (13). We do this separately for the two cases above and then arrive at a uniform lower bound. Let k^* be the largest k such that $|S_k| < n$. For the first case, when $k \leq k^*$,

$$\begin{aligned}\binom{n}{r} \Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\} &\geq \frac{3}{4} \binom{\frac{1}{4} \frac{\binom{k}{d}}{\binom{rd}{d}}}{r} && (15) \\ &\geq \frac{3}{4} \frac{1}{(4r)^r} \left(\frac{\binom{k}{d}}{\binom{rd}{d}} \right)^r && (16) \\ &\geq \frac{3}{4} \frac{1}{(4r)^r} \left(\frac{\left(\frac{k}{d}\right)^d}{\left(\frac{rde}{d}\right)^d} \right)^r \\ &= \frac{3}{4} \frac{k^{rd}}{(4r)^r (erd)^{rd}}. && (17)\end{aligned}$$

For the second case, we claim that it is sufficient to only consider codeword lengths up to a particular k_{max} in our calculation of the expected length. To see this, observe that the probability of encoding by codewords of length $k^* + 1$ is given by

$$\begin{aligned}\Pr\{\mathcal{E}_{k^*+1,1} \cap \mathcal{E}_{k^*+1,2}\} &\geq 1 - n \frac{\binom{rd}{d}}{\binom{k^*+1}{d}} \\ &\geq 3/4. && (18)\end{aligned}$$

Let us now consider the codeword lengths $k^* + 1, k^* + 2, \ldots, k^* + t$. Since $\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\}$ increases with k, an application of (18) gives us the following upper bound on the probability that x^n is not encoded by codewords of any of these lengths.

$$\Pr\{x^n \text{ not encoded by any codewords of length } k^* + 1, k^* + 2, \ldots, k^* + t\} \leq \left(\frac{1}{4}\right)^t.$$

Since the total number of source sequences is $\binom{n}{r}$, we can see that setting $t = \frac{\log \binom{n}{r}}{2}$ is sufficient to ensure that $\binom{n}{r} \left(\frac{1}{4}\right)^t < 1$. In other words, there exists a code that ensures that all source sequences in $\binom{[n]}{r}$ are successfully encoded into codewords of length between $k^* + 1$ and $k^* + t$. Therefore, it is sufficient for us to restrict our attention to k such that $k \leq k_{\mathsf{max}}$, with k_{max} defined as

$$k_{\mathsf{max}} \triangleq k^* + \frac{\log \binom{n}{r}}{2}. \qquad (19)$$

We know, however, from (2), that $k_{\mathsf{max}} \geq \log \binom{n}{r} - \log \log \binom{n}{r} - \log e$. For $\log \binom{n}{r} \geq 12$, it is straightforward to combine this fact with (19) to obtain

$$k_{\mathsf{max}} \leq 8k^*. \qquad (20)$$

Bounding for the second case, we now have

$$\begin{aligned}\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\} &\geq \Pr\{\mathcal{E}_{k^*,1} \cap \mathcal{E}_{k^*,2}\}, \text{ for } k^* < k \leq k_{\mathsf{max}} && (21) \\ &\geq \frac{3}{4} \frac{(k^*)^{rd}}{(4r)^r (erd)^{rd} \binom{n}{r}} && (22) \\ &\geq \frac{3}{4} \frac{(k/8)^{rd}}{(4r)^r (erd)^{rd} \binom{n}{r}}, && (23)\end{aligned}$$

where (21) follows from the fact that $\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\}$ increases with k, (22) follows from (17), and (23) from (20). Hence, from (17) and (23), we have the uniform lower bound

$$\Pr\{\mathcal{E}_{k,1} \cap \mathcal{E}_{k,2}\} \geq \mathsf{C}_{r,d} k^{rd}, \quad \forall\ rd + 1 \leq k \leq k_{\mathsf{max}}, \qquad (24)$$

where we have defined $\mathsf{C}_{r,d} \triangleq \left(\frac{4}{3} \binom{n}{r} (4r)^r (8erd)^{rd}\right)^{-1}$ for convenience.

Now, using the inequality $(1 - x) \leq e^{-x}$, we can upper bound (13) with (24) as

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}} \left[\ell(\mathsf{c}(x^n))\right] \\ &\leq \sum_{k=rd+1}^{k_{\mathsf{max}}} \mathsf{C}_{r,d} k^{rd+1} \exp\left(- \sum_{j=rd+1}^{k-1} \mathsf{C}_{r,d} j^{rd}\right) \\ &\leq \sum_{k \geq rd+1} \mathsf{C}_{r,d} k^{rd+1} \exp\left(-\mathsf{C}_{r,d} \int_{rd+1}^{k-1} z^{rd} dz\right) \\ &= \sum_{k \geq rd+1} \Bigg[\mathsf{C}_{r,d} k^{rd+1} \\ &\qquad \times \exp\left(-\mathsf{C}_{r,d} \frac{\left((k-1)^{rd+1} - (rd+1)^{rd+1}\right)}{rd+1}\right)\Bigg] \\ &= \exp\left(\mathsf{C}_{r,d} (rd+1)^{rd}\right) \\ &\qquad \times \sum_{k \geq rd+1} \mathsf{C}_{r,d} k^{rd+1} \exp\left(-\mathsf{C}_{r,d} \frac{(k-1)^{rd+1}}{rd+1}\right) \\ &\leq \exp\left(\mathsf{C}_{r,d} (rd+1)^{rd} + \frac{rd+1}{rd}\right) \\ &\qquad \times \sum_{k \geq rd+1} \mathsf{C}_{r,d} (k-1)^{rd+1} \exp\left(-\mathsf{C}_{r,d} \frac{(k-1)^{rd+1}}{rd+1}\right) \\ &= \exp\left(\mathsf{C}_{r,d} (rd+1)^{rd} + \frac{rd+1}{rd}\right) (rd+1) \\ &\qquad \times \sum_{k \geq rd} \mathsf{C}_{r,d} \frac{k^{rd+1}}{rd+1} \exp\left(-\mathsf{C}_{r,d} \frac{k^{rd+1}}{rd+1}\right). && (25)\end{aligned}$$

Since the function ue^{-u} is monotone increasing on $(0, 1)$ and monotone decreasing on $(1, \infty)$ with a maximum of $1/e$, we can bound the sum in (25) as

$$\sum_{k=rd}^{\infty} \mathsf{C}_{r,d} \frac{k^{rd+1}}{rd+1} \exp\left(-\mathsf{C}_{r,d} \frac{k^{rd+1}}{rd+1}\right) \leq 2/e + \int_0^{\infty} \mathsf{C}_{r,d} \frac{z^{rd+1}}{rd+1} \exp\left(-\mathsf{C}_{r,d} \frac{z^{rd+1}}{rd+1}\right) dz. \quad (26)$$

The integral in (26) can be bounded by the following lemma, the proof of which is postponed to the appendix.

Lemma 2:

$$(rd+1) \int_0^{\infty} \mathsf{C}_{r,d} \frac{z^{rd+1}}{rd+1} \exp\left(-\mathsf{C}_{r,d} \frac{z^{rd+1}}{rd+1}\right) dz \leq \left(\frac{rd+1}{\mathsf{C}_{r,d}}\right)^{1/(rd+1)}.$$

To finish the proof, we use Lemma 2 and the integral upper bound on (25) to conclude

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\left[\ell(\mathsf{c}(x^n))\right] \\
&\leq 2(rd+1) \exp\left(\mathsf{C}_{r,d}(rd+1)^{rd} + \frac{1}{rd}\right) \\
&\quad + \exp\left(\mathsf{C}_{r,d}(rd+1)^{rd} + \frac{rd+1}{rd}\right)\left(\frac{rd+1}{\mathsf{C}_{r,d}}\right)^{1/(rd+1)} \\
&\leq 2(rd+1)e \cdot \exp\left(\frac{3}{128n}\right) \\
&\quad + e(rd+1)\exp\left(\frac{3}{128n}\right) \cdot \left(r^r \binom{n}{r}\right)^{1/(rd+1)} \\
&\quad \times \left[\exp\left(\frac{1}{rd}\right) \cdot (8e)^{rd/(rd+1)} \left(\frac{4 \cdot 4^r}{3}\right)^{1/(rd+1)}\right] \quad (27) \\
&\leq 250(rd+1)\left(r^r \binom{n}{r}\right)^{1/(rd+1)}, \quad (28)
\end{aligned}$$

where (27) comes from substituting the definition of $\mathsf{C}_{r,d}$, and (28) comes from optimizing the quantity in square brackets - which increases in r and decreases in d - and assuming $n \geq 1$.

Since all sequences in $\binom{[n]}{r}$ are equally probable, linearity of expectation ensures the existence of a code c which satisfies (5) as desired.

Remark 1: Notice that the constant factor in (28) *differs substantially from that of the lower bound* (3). *However, the bounds are tight for the case of $r = d = 1$ as shown in Section III-A.1. In addition, for $d = 1$ and arbitrary r, the analysis can be tightened to improve the constant factor to $\Gamma\left(\frac{r+2}{r+1}\right)$. Based on this, we conjecture that the achievability scheme proposed in the proof of Theorem 2 can yield a multiplicative factor as small as $\Gamma\left(\frac{rd+2}{rd+1}\right)$, which is strictly less than 1, at the expense of more careful intermediate bounds.* ■

Proof of Theorem 3: Suppose $\{\mathsf{c}_n\}$ is a competitively optimal sequence of (r_n, d_n, n)-locally decodable codes, and define ϵ_n according to

$$r_n d_n + 1 = \epsilon_n \log \binom{n}{r_n}.$$

Then Theorem 1 and the definition of competitive optimality imply that there is some constant K for which

$$\begin{aligned}
K &\geq \frac{\mathbb{E}[\ell(\mathsf{c}_n(X^n))]}{\log \binom{n}{r_n}} \\
&\geq \epsilon_n \left(\binom{n}{r_n}^{1/(\epsilon_n \log \binom{n}{r_n})} - 1\right) \\
&= \epsilon_n \left(2^{1/\epsilon_n} - 1\right)
\end{aligned}$$

for all n sufficiently large. Since $\epsilon(2^{1/\epsilon} - 1) \nearrow \infty$ as $\epsilon \searrow 0$, this implies that there is a constant $K' > 0$ such that $\epsilon_n \geq K'$ for all n sufficiently large, proving the claim. ■

Proof of Theorem 4: Theorem 4 is a consequence of the following general result which provides an explicit expression for the $o(1)$ term.

Lemma 3: There exists a sequence of (r_n, d_n, n) codes with expected codeword length satisfying (5) *for which $D \leq 1 + \frac{1}{C-1} + \frac{d(Ce)^d + rd}{n}$ bits.*

Proof: We use the scheme from the proof of Theorem 2, with the encoding procedure modified such that the source sequence is stored as is when $k = n$. Note that the code still remains locally decodable, since the sequences that are stored without compression are 1-locally decodable. This ensures that the maximum codeword length is upper bounded at n.

The decoding procedure is modified as follows.

(D1) If $j \notin S_{\ell(\mathsf{c}(x^n))}$, declare $x_j = 0$; else

(D2) If $j \in S_{\ell(\mathsf{c}(x^n))}$, probe bits in $\cup_{i \in T_{j,k(x^n)}} c_i$ uniformly at random with replacement.[7] Stop when the first 0 is seen, and declare $x_j = 0$, else declare $x_j = 1$.

Intuitively, it should be obvious that this scheme would probe a much smaller number of bits than the scheme of Theorem 2 on average, since we "exit" after making just a few probes whenever the bit being queried is a zero. We now show that this is indeed the case.

Let $\mathcal{C}$ denote the (random) codebook. For fixed codeword length k, we denote the average number of queries made over all codewords of that length by

$$\begin{aligned}
\mathsf{D}_k(\mathcal{C}) &\triangleq \mathbb{E}_{x^n:\ell(\mathsf{c}(x^n))=k} \mathbb{E}_{i \in [n]} d(i, x^n) \\
&= \frac{1}{\#\{x^n_{(i)} : \ell(\mathsf{c}(x^n_{(i)})) = k\}} \sum_{x^n:\ell(\mathsf{c}(x^n))=k} \frac{1}{n} \sum_{i \in [n]} d(i, x^n).
\end{aligned}$$

Let us analyze the decoding of bits when the codeword length is k. Out of the n bits to be decoded, for the $n - \min\left(n, \frac{r}{r+1} \frac{\binom{k}{d}}{\binom{rd}{d}}\right)$ bits corresponding to condition (D1) of the decoding rule, we declare $x_j = 0$ after simply looking at $\ell(\mathsf{c}(x^n))$ (i.e. without a single probe). For the other $\min\left(n, \frac{r}{r+1} \frac{\binom{k}{d}}{\binom{rd}{d}}\right) - r$ bits that are zeroes corresponding to condition (D2), we make a number of probes that is a truncated geometric random variable with support $[d]$, and success probability at least $\frac{k-rd}{k}$, since the number of ones in the codeword is upper bounded by rd. The expected number of probes to determine these bits is therefore upper bounded by $\min(d, \frac{k}{k-rd})$. For the remaining r bits that are ones, we make

[7] These results carry over to probes without replacement.

d probes to determine them. We can therefore conclude that over the random codebook,

$$\mathbb{E}_{\mathcal{C}}[\mathsf{D}_k(\mathcal{C})] \le \frac{1}{n}\Bigg(\min\Bigg(n, \frac{r}{r+1}\frac{\binom{k}{d}}{\binom{rd}{d}}\Bigg)\min\Bigg(d, \frac{k}{k-rd}\Bigg)+rd\Bigg).$$

Taking the expectation of this over the probability distribution of $\ell(\mathsf{c}(x^n))$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}[\mathsf{D}(\mathcal{C})] &= \sum_{k \ge rd+1} \Pr\{\ell(\mathsf{c}(x^n)) = k\} \cdot \mathbb{E}_{\mathcal{C}}[\mathsf{D}_k(\mathcal{C})] \\ &\le \frac{rd}{n} + \frac{1}{n}\sum_{k\ge rd+1} \Pr\{\ell(\mathsf{c}(x^n)) = k\} \\ &\quad \times \Bigg(\frac{r}{r+1}\frac{\binom{k}{d}}{\binom{rd}{d}} \min\Bigg(d, \frac{k}{k-rd}\Bigg)\Bigg), \qquad (29)\end{aligned}$$

where we have implicitly assumed that we sum over k such that $\frac{r}{r+1}\frac{\binom{k}{d}}{\binom{rd}{d}} \le n$.

Splitting the sum in (29) into two parts for some fixed $C > 1$, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}}[\mathsf{D}(\mathcal{C})] \\ &\le \frac{rd}{n} + \frac{1}{n}\sum_{k=rd+1}^{Crd} \Pr\{\ell(\mathsf{c}(x^n)) = k\} \cdot \Bigg(\frac{r}{r+1}\frac{\binom{k}{d}}{\binom{rd}{d}} \cdot d\Bigg) \\ &\quad + \frac{1}{n}\sum_{k \ge Crd+1} \Pr\{\ell(\mathsf{c}(x^n)) = k\} \cdot \Bigg(\frac{r}{r+1}\frac{\binom{k}{d}}{\binom{rd}{d}} \cdot \frac{C}{C-1}\Bigg), \\ &\qquad\qquad (30)\end{aligned}$$

since $\frac{k}{k-rd} < \frac{C}{C-1}$ for $k \ge Crd+1$. Bounding each of the two sums, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}[\mathsf{D}(\mathcal{C})] &\le \frac{rd}{n} + \frac{1}{n}d \cdot \Bigg(\frac{r}{r+1}\frac{(\frac{e\cdot Crd}{d})^d}{(\frac{rd}{d})^d}\Bigg) + \frac{C}{C-1} \\ &\le \frac{d(Ce)^d + rd}{n} + \frac{C}{C-1},\end{aligned}$$

where we have used the fact that $\frac{r}{r+1}\frac{\binom{k}{d}}{\binom{rd}{d}} \le n$, since the codeword is never longer than n bits. As before, linearity of expectation ensures the existence of a sequence of (n, r, d) codes in the random ensemble with $\mathsf{D} \le \frac{d(Ce)^d+rd}{n} + \frac{C}{C-1}$. ■

Proof of Theorem 5: Recall the definition of an α-optimal code, as being one which has expected codeword length within a constant factor α of the lower bound (3).

Also recall the definition of average probe depth to recover 1s of the source D^1, which is restated below for convenience.

$$\mathsf{D}^1 = \mathbb{E}_{x^n}\mathbb{E}_{i\in\mathsf{supp}(x^n)} d(i, x^n).$$

Now define the quantity D^1_k as being the average probe depth D^1 conditioned on the codeword length being k. In other words,

$$\begin{aligned}\mathsf{D}^1_k &\triangleq \mathbb{E}_{x^n:\ell(\mathsf{c}(x^n))=k}\mathbb{E}_{i\in\mathsf{supp}(x^n)} d(i, x^n) \\ &= \frac{1}{\#\{x^n_{(i)} : \ell(\mathsf{c}(x^n_{(i)})) = k\}} \sum_{x^n:\ell(\mathsf{c}(x^n))=k} \frac{1}{r}\sum_{i\in\mathsf{supp}(x^n)} d(i, x^n).\end{aligned}$$

In order to prove Theorem 5, we first show that it is sufficient to prove the following proposition.

Proposition 2: Suppose $r_n d_n = o\left(\log\binom{n}{r}\right)$. Then for any α-optimal code, there exists some $S \subset \mathbb{N}$ such that

1) $\mathsf{D}^1_k \ge d - \frac{C}{r}\ \forall\ k \in S$, and

2) $\frac{1}{\binom{n}{r}}\sum_{k\in S}\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \ge C'$,

for some constants C, C' that depend on α.

It is easy to see that Proposition 2 proves Theorem 5, since

$$\begin{aligned}\mathsf{D}^1 &= \frac{1}{\binom{n}{r}}\sum_k \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}\mathsf{D}^1_k \\ &\ge \frac{1}{\binom{n}{r}}\sum_{k\in S} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}\mathsf{D}^1_k \\ &\ge \Bigg(d - \frac{C}{r}\Bigg)\frac{1}{\binom{n}{r}}\sum_{k\in S} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \\ &\ge C'\Bigg(d - \frac{C}{r}\Bigg).\end{aligned}$$

We have therefore established that it is sufficient to prove Proposition 2. In order to do so, we need the following key lemma, which relates D^1_k to $\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}$ for large enough k. Note its similarity to (8).

Lemma 4: For any codeword length $k \ge 2rd + 4$, the following holds:

$$2^{r\mathsf{D}^1_k+2}\binom{k}{r\mathsf{D}^1_k + 2} \ge \frac{1}{2}\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}.$$

proof: Define the transcript T^k_i as in (7). Let $m^k_t = \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k, |T^k_i| = t\}$. In words, m^k_t is the number of source sequences that are encoded into codewords of length k and have transcripts of size t. Note that by definition,

$$\sum_{t=0}^{rd} t m^k_t \le r\mathsf{D}^1_k \cdot \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\},$$

and

$$\sum_{t=0}^{rd} m^k_t = \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}.$$

Also, we know that m^k_t is bounded by the number of unique T^k_is that can be created. This constraint can be expressed as:

$$m^k_t \le 2^t\binom{k}{t}.$$

From these constraints, we see that the average probe depth $D^1_k \ge p^*$, where p^* is the solution to the following linear program:

$$\begin{aligned}p^* = \min_{a_t, p}\ & p \\ \text{s. t. } & \sum_{t=0}^{rd} t a_t \le rp \cdot \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \\ & \sum_{t=0}^{rd} a_t = \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \\ & a_t \le 2^t\binom{k}{t} \quad \forall\ t = \{1, 2, \ldots rd\}. \qquad \text{(P1)}\end{aligned}$$

The solution to (P1) can be found by a greedy argument, as follows. We want

$$\sum_{t=1}^{\hat{t}-1} t 2^t \binom{k}{t} + \hat{t} a_{\hat{t}} = rp \cdot \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}, \tag{31}$$

where $\hat{t}$ is the smallest integer satisfying

$$\sum_{t=1}^{\hat{t}} 2^t \binom{k}{t} \geq \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}. \tag{32}$$

Here, $a_{\hat{t}} = \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} - \sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t}$.

Since $\frac{2^{i+1}\binom{s}{i+1}}{2^i \binom{s}{i}} \geq 2$ for $i < s/2$ and $\hat{t} \leq rd$, we have

$$\frac{\sum_{t=1}^{\hat{t}-1} t 2^t \binom{k}{t}}{\sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t}} \geq (\hat{t} - 2) \tag{33}$$

for $k > 2rd - 2$. We therefore have, from (31) and (33), that

$$\begin{aligned} rp^* &\geq \frac{(\hat{t}-2)\sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t} + \hat{t} a_{\hat{t}}}{\sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t} + a_{\hat{t}}} \\ &= \hat{t} - 2 \frac{\sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t}}{\sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t} + a_{\hat{t}}} \\ &\geq \hat{t} - 2 \end{aligned} \tag{34}$$

We also know that $2^{\hat{t}} \binom{k}{\hat{t}} \geq \sum_{t=1}^{\hat{t}-1} 2^t \binom{k}{t}$, and so by (32),

$$2^{\hat{t}} \binom{k}{\hat{t}} \geq \frac{1}{2} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}.$$

Equation (34) gives us that

$$2^{rp^*+2} \binom{k}{rp^*+2} \geq \frac{1}{2} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\},$$

which proves the lemma since $2^i \binom{s}{i}$ is an increasing function of i for $i < s/2$, and $\mathsf{D}^1_k \leq d \ \forall\ k$. □

Our next step is to lower bound $\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}$ with some function of d for α-optimal codes. Let us begin with some definitions. Define $p_{\mathsf{max}}(k) = \frac{\max_{\upsilon \leq rd} \binom{2k}{\upsilon}}{\binom{n}{r}}$. For any (r, d, n) locally decodable code, define a distribution $p : p(k) = \frac{\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\}}{\binom{n}{r}}$. By the argument used in the proof of Theorem 1, $p(k) \leq p_{\mathsf{max}}(k) \ \forall\ k$; call such distributions *valid distributions*.

The following technical lemma proves a key property of such distributions. The proof is postponed to the Appendix D.

Lemma 5: Let p be a valid distribution corresponding to an α-optimal (r, d, n)-locally decodable code. Then there exist constants $\epsilon(\alpha), \delta(\alpha)$ independent of r, d, n, and $S \subseteq \mathbb{N}$ such that

1) $\alpha^{rd} \frac{p(k)}{p_{\mathsf{max}}(k)} \geq \epsilon(\alpha) \ \forall\ k \in S$, and

2) $\sum_{k \in S} p(k) \geq \delta(\alpha)$.

Abbreviating $\epsilon(\alpha), \delta(\alpha)$ by ϵ, δ and restating Lemma 5 in words: for any scheme that has expected length within a constant factor of (3), there is a set S such that $\frac{1}{\binom{n}{r}} \sum_{k \in S} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \geq \delta$, and for which $\#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \geq \frac{\epsilon}{\alpha^{rd}} \max_{\upsilon \leq rd} \binom{2k}{\upsilon} \ \forall\ k \in S$. Combining this with Lemma 4 for $k \geq 2rd + 4$, we obtain

$$\begin{aligned} \binom{2k}{r\mathsf{D}^1_k + 2} &\geq 2^{r\mathsf{D}^1_k + 2} \binom{k}{r\mathsf{D}^1_k + 2} \\ &\geq \frac{1}{2} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \\ &\geq \frac{\epsilon}{2\alpha^{rd}} \max_{\upsilon \leq rd} \binom{2k}{\upsilon} \\ &= \frac{\epsilon}{2\alpha^{rd}} \binom{2k}{rd}. \end{aligned} \tag{35}$$

Rearranging (35), we obtain

$$\frac{2\alpha^{rd}}{\epsilon} \geq \left(\frac{2k - rd + 1}{rd} \right)^{rd - r\mathsf{D}^1_k - 2},$$

which for $k \geq (\alpha + 1/2)rd$ yields

$$\log\left(\frac{2}{\epsilon}\right) \geq rd - r\mathsf{D}^1_k - 2.$$

Therefore, for each $k \in S \cap \{(\alpha + 1/2)rd, \ldots\} \cap \{2rd + 4, 2\,rd + 5, \ldots\}$, we have

$$\mathsf{D}^1_k \geq d - \tfrac{1}{r}\left(\log\left(\tfrac{2}{\epsilon}\right) + 2\right).$$

Defining $M' \triangleq \min((\alpha + 1/2)rd, 2rd + 4)$, the only remaining step to prove Proposition 2 is to show that

$$\frac{1}{\binom{n}{r}} \sum_{k=1}^{M'} \#\{i : \ell(\mathsf{c}(x^n_{(i)})) = k\} \to 0 \quad \text{as } n \to \infty.$$

This is equivalent to showing that

$$\lim_{n \to \infty} \sum_{k \leq M'} p(k) = 0. \tag{36}$$

Since $p(k) \leq p_{\mathsf{max}}(k)$, an application of Lemma 1 with $M \leftarrow M'$ proves (36) for $\binom{n}{r}^{\frac{1}{(rd+1)}} = \omega(1)$.

Proposition 2, and with it Theorem 5, are therefore proved. ■

Proof of Theorem 6: We use a set of $p = \frac{1}{d} \log \binom{n}{r}$ *schemes* $s_1, s_2, \ldots s_p$, each of which uses a data structure of blocklength $\ell = 3rdn^{1/d}$ to encode a subset of $\binom{[n]}{r}$. Information about which scheme is used in the encoding will be stored in the header, and so a header length of $\log p$ bits is necessary.

E. Codebook Construction

For each $k \in [p]$ and $j \in [n]$, choose a subset $T_{j,k} \subseteq [\ell]$ of size d independently and uniformly from all such subsets. All subsets (i.e., the codebook) are made available to both encoder and decoder.

For a sequence $x^n \in \binom{[n]}{r}$, let $k(x^n)$ denote the smallest integer k such that the following condition holds:

- $T_{j,k} \not\subseteq \cup_{i \in \mathsf{supp}(x^n)} T_{i,k}$ for all $j \in [n] \setminus \mathsf{supp}(x^n)$.

In words, $k(x^n)$ is the first scheme among $s_1 \ldots s_p$ using which x^n can be encoded, as follows.

F. Encoding Procedure

A sequence $x^n \in \binom{[n]}{r}$ is encoded to a codeword $\mathsf{c}(x^n)$ of length ℓ satisfying

$$\mathsf{supp}(\mathsf{c}(x^n)) = \cup_{i \in \mathsf{supp}(x^n)} T_{i,k(x^n)},$$

with scheme number given by the binary representation of $k(x^n)$.

In other words, x^n is encoded to a vector of length ℓ, which has 1's in all positions $j \in T_{i,k(x^n)}$ if and only if $x_i = 1$, and a header is prefixed to it denoting the number of the scheme using which it was encoded.

G. Decoding Procedure

On observing the scheme number $k(x^n)$, determine bit x_j trivially as follows:

Declare $x_j = \wedge_{i \in T_{j,k(x^n)}} c_i$, where '$\wedge$' denotes binary AND.

By the nature of the codebook construction, it is clear that the decoder (i) will never make an error; and (ii) satisfies the non-adaptive d-local decodability constraint.

H. Performance Analysis

We must now show that our choices of $\ell = (2+\epsilon) r d n^{1/d}$ and $p = \frac{1}{d} \log \binom{n}{r}$ allow for a correct encoding.

Let us start by looking at the number of source sequences that are encoded by scheme s_1. Consider an arbitrary source sequence x^n. Define the bad event

$$\mathcal{E}_j = \{x_j = 0, T_{j,1} \subseteq \cup_{i \in \mathsf{supp}(x^n)} T_{i.1}\}.$$

Since $T_{j,1}$s are picked independently and uniformly,

$$\Pr\{\mathcal{E}_j\} \leq \left(\frac{rd}{\ell}\right)^d \quad \forall\ j \in [n] \setminus \mathsf{supp}(x^n).$$

Now x^n is not encoded in scheme s_1 if and only if $\mathcal{E}_j$ occurs for some $j \in [n] \setminus \mathsf{supp}(x^n)$. Therefore,

$$\begin{aligned}\Pr\{x^n \text{ is not encoded using } s_1\} &= \Pr\{\cup_{j \in [n] \setminus \mathsf{supp}(x^n)} \mathcal{E}_j\} \\ &\leq (n-r)\left(\frac{rd}{\ell}\right)^d.\end{aligned}$$

It is therefore clear by linearity of expectation that the expected number of source sequences that are available for encoding by $s_2, s_3, \ldots, s_p$ is given by

$$\mathbb{E}\left[\#\{i : x^n_{(i)} \text{ is not encoded by } s_1\}\right] \leq \binom{n}{r}(n-r)\left(\frac{rd}{\ell}\right)^d.$$

Since the randomness used in the creation of $T_{j,k}$ for each scheme k is independent of the others, we can use the argument above to see that after p schemes, the expected number of source sequences that are left to be decoded is

$$\begin{aligned}&\mathbb{E}\left[\#\{i : x^n_{(i)} \text{ not encoded by } s_1, s_2, \ldots, s_p\}\right] \\ &\qquad \leq \binom{n}{r}\left[(n-r)\left(\frac{rd}{\ell}\right)^d\right]^p.\end{aligned}$$

It is straightforward to verify that our choices of ℓ and p ensure that this quantity is less than 1, which ensures the existence of a deterministic set of schemes $s_1, \ldots, s_p$ that encode all $\binom{n}{r}$ source sequences. ■

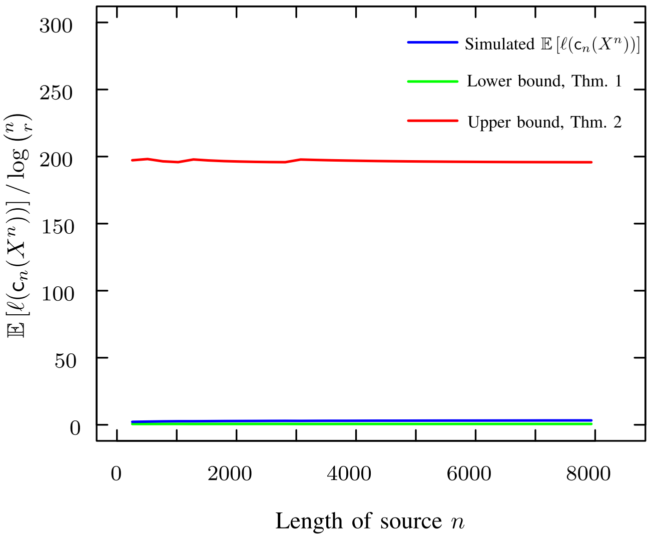

Fig. 2. Plot of the simulated expected codeword length along with the upper and lower bounds for a competitively optimal sequence of codes, with $d = \log n$, $r = 5$. Notice the normalization by the entropy of the source, and the fact that the bounds only differ by a constant factor as predicted by Corollary 2.

V. Numerical Experiments

Fixing an r-sparse codeword x^n, we simulated the randomized scheme of Theorem 2, noting the length of the codeword at which x^n first satisfied the conditions for encoding. We averaged this over 100 trials, and plotted our results for a sequence of competitively optimal codes with fixed $r = 5$ and by varying n with $d = \log n$.

Figure 2 confirms the scaling suggested by corollary 2. Rather surprisingly, the lower bound of Theorem 1 appears to be close to the empirical performance of our encoding scheme. The plot also reveals a substantial gap in the constant factor between the upper bound predicted by Theorem 2 and the simulation, suggesting a weakness in the analysis of our scheme.

In particular, we have sacrificed constant factors in going from (15) to (16) and again in (24), in order to get uniform bounds which hold for all n, r, d. As the plot reveals, knowledge of how r and d scale can be used to tighten the analysis and improve the constant factor substantially.

VI. Concluding Remarks

We provided bounds for the blocklength scaling behaviour of (r, d, n) locally-decodable codes that are tight up to constant factors for many regimes of r, d, and n. Determining the tight constant in these bounds is an open problem. We also showed that in contrast to the fixed blocklength setting (cf. [14]), adaptivity of probes provides no essential advantage in our setting of variable length source coding. An interesting open question is the extension of these results to the case when the source distribution is non-uniform, in which case variable blocklength codes are clearly preferable to fixed blocklength ones. In conclusion, we mention two variations on our main results:

A. Compression With Block Errors

In [21], Makhdoumi *et al.* allow for vanishing block-error probability in decoding. Although we only considered error-free encodings, the proof of Theorem 1 readily extends to incorporate block-error probability as follows: Letting $\hat{x}^n$ denote the decoder's estimate of the sequence x^n given codeword $\mathsf{c}(x^n)$, the block error rate is defined to be $\Pr\{X^n \neq \hat{X}^n\}$. Now, Theorem 1 continues to hold for any (r, d, n) code with block-error rate ε by simply replacing the quantity $\binom{n}{r}$ with $(1-\varepsilon)\binom{n}{r}$. Indeed, this follows by considering only those sequences that are correctly decoded and making the same substitution in (10) in the proof of Theorem 1.

B. Connection to Communication Complexity

It is known that the bit-probe model has applications to asymmetric communication complexity [26]. To draw an analogous connection to our setting, consider an asymmetric communication complexity model [26] in which Alice (the user) has $i \in [n]$, Bob (the server) has $S \subset [n]$ of size r, and they wish to compute the membership function

$$f(i, S) = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

We now enforce that the function f must be computed under a SpeedLimit paradigm, which proceeds as follows. Communication starts with Bob sending a *speed limit* message to Alice consisting of some z bits, which limits the length of any of her messages to z bits. Bob's subsequent messages consist of 1 bit. After the initial round, Alice and Bob communicate over d rounds[8] to evaluate f. The setting arises in practice where a server imposes upload bandwidth limits on users it serves (e.g., to maintain quality or fairness of service).

Note that our scheme in Theorem 2 provides a communication protocol to compute f under the SpeedLimit paradigm. Bob is essentially given a source sequence x^n, which he stores as $c(x^n)$. Alice is given the index i of the source bit that must be decoded, and must do so by making queries to Bob. Bob begins by sending $\ell(\mathsf{c}(x^n))$ to Alice, using $\log \ell(\mathsf{c}(x^n))$ bits. Alice then sends messages m_j, $j \in [d]$ of $\log \ell(c(x^n))$ bits each. In response to message m_j, Bob sends back $c_{m_j}(x^n)$. Alice then announces f to be the AND of the d bits she has received from Bob. Therefore, from Theorem 2:

Corollary 3: There exists a deterministic communication protocol for computing the function f as in (37) *under the* SpeedLimit *paradigm for which the speed limit z and number of communication rounds d satisfy*

$$\mathbb{E}\left[2^z\right] \leq 250(rd+1)\left(r^r \binom{n}{r}\right)^{1/(rd+1)}.$$

Appendix
Technical Proof Details

In this section, we provide further details for the proofs given in Section III.

A. Details for Proof of Theorem 1

The LYM inequality used in proving the lower bound (3) is given below for convenience.

Lemma 6 (LYM Inequality [25]): Let U be a u-element set, let $\mathcal{A}$ be a family of subsets of U such that no set in $\mathcal{A}$ is a subset of another set in $\mathcal{A}$, and let a_m denote the number of sets of size m in $\mathcal{A}$. Then

$$\sum_{m=0}^{u} \frac{a_m}{\binom{u}{m}} \leq 1. \tag{38}$$

The second key inequality was Lemma 1, which is restated below for convenience.

Lemma 1: For all $M, \upsilon \geq 1$,

$$\sum_{k=1}^{M} \max_{i \leq \upsilon} \binom{2k}{i} \leq (\upsilon+1)^{\frac{1}{2}} 2^{\upsilon} \frac{\left(M+2+\frac{\upsilon+1}{2e}\right)^{\upsilon+1}}{(\upsilon+1)!}. \tag{39}$$

Proof of Lemma 1: We begin the proof by splitting the bound into two cases:

Proposition 3: For $\upsilon \geq 1$

$$\sum_{k=1}^{M} \max_{i \leq \upsilon} \binom{2k}{i} \leq \begin{cases} \frac{3}{2}\binom{2M}{M} & \text{for } 1 \leq M \leq \upsilon \\ 2^{\upsilon} \frac{(M+2)^{\upsilon+1}}{(\upsilon+1)!} & \text{for } M \geq \upsilon + 1. \end{cases}$$

proof: Note that if $M \leq \upsilon$, then

$$\sum_{k=1}^{M} \max_{i \leq \upsilon} \binom{2k}{i} = \sum_{k=1}^{M} \binom{2k}{k}.$$

We prove by induction on M that $\sum_{k=1}^{M} \binom{2k}{k} \leq \frac{3}{2}\binom{2M}{M}$. The base case $M = 1$ is trivial, so by the inductive hypothesis, we have

$$\sum_{k=1}^{M+1} \binom{2k}{k} \leq \binom{2(M+1)}{M+1} + \frac{3}{2}\binom{2M}{M}.$$

However, we can write

$$\binom{2(M+1)}{M+1} = \frac{(2M+2)(2M+1)}{(M+1)^2}\binom{2M}{M} \geq 3\binom{2M}{M},$$

completing the proof of the first claim.

Next, if $M \geq \upsilon + 1$, then

$$\begin{aligned} \sum_{k=1}^{M} \max_{i \leq \upsilon} \binom{2k}{i} &= \sum_{k=1}^{\upsilon} \binom{2k}{k} + \sum_{k=\upsilon+1}^{M} \binom{2k}{\upsilon} \\ &\leq \sum_{k=\upsilon}^{M+1} \binom{2k}{\upsilon} \\ &\leq \sum_{k=\upsilon}^{M+1} \frac{(2k)^{\upsilon}}{\upsilon!} \\ &\leq \frac{2^{\upsilon}}{\upsilon!} \int_0^{M+2} z^{\upsilon} dz \\ &\leq 2^{\upsilon} \frac{(M+2)^{\upsilon+1}}{(\upsilon+1)!}, \end{aligned}$$

establishing the second claim. □

[8] To be consistent with the rest of the paper, a communication round consists of one message by Alice and a response by Bob.

Recalling the crude upper bound $\binom{2M}{M} \le 4^M$, Lemma 1 follows from the inequality

$$(3(v+1)!)^{1/(v+1)} 4^{M/(v+1)} \le (v+1)^{\frac{1}{2(v+1)}} 2\left(M+2+\frac{v+1}{2e}\right) \tag{40}$$

for $1 \le M \le v$ and Proposition 3. Using the well-known bound $n! \le en^{n+\frac{1}{2}}e^{-n}$, we have

$$\begin{aligned}&(3(v+1)!)^{1/(v+1)}4^{M/(v+1)}\\ &\qquad \le (3e)^{1/(v+1)}(v+1)(v+1)^{\frac{1}{2(v+1)}}4^{M/(v+1)}e^{-1}.\end{aligned}$$

Thus, (40) is implied by the following:

$$(3e)^{1/(v+1)}(v+1)4^{M/(v+1)} \le 2e\left(M+2+\frac{v+1}{2e}\right). \tag{41}$$

Since the LHS in (41) is a convex function in M, we can show (40) holds for $1 \le M \le v$ by verifying it at the points $M = 0$ and $M = v+1$. It is straightforward to check that this is the case, completing the proof. Indeed, for $M = 0$, (41) reduces to

$$(3e)^{1/(v+1)} - \frac{4e}{v+1} \le 1,$$

which holds since

$$\max_{0 \le x \le 1/2}\left\{(3e)^x - 4ex\right\} = 1.$$

Similarly, for $M = v+1$, (41) reduces to

$$4(3e)^{1/(v+1)} - \frac{4e}{v+1} \le 1+2e,$$

which holds since

$$\begin{aligned}\max_{0 \le x \le 1/2}\left\{4(3e)^x - 4ex\right\} &= 4\left(\sqrt{3e}-\frac{e}{2}\right)\\ &< 1+2e.\end{aligned}$$

■

B. Details for Proof of Theorem 2

In this section, we provide the proof of Lemma 2, which is restated below for convenience.

Lemma 2:

$$\begin{aligned}(rd+1)\int_0^\infty \mathsf{C}_{r,d}\frac{z^{rd+1}}{rd+1}\exp\left(-\mathsf{C}_{r,d}\frac{z^{rd+1}}{rd+1}\right)dz\\ \le \left(\frac{rd+1}{\mathsf{C}_{r,d}}\right)^{1/(rd+1)}.\end{aligned}$$

Proof: Abbreviating $C \triangleq \mathsf{C}_{r,d}$, consider the change of variables $z = \left(\frac{u(rd+1)}{C}\right)^{1/(rd+1)}$. Then,

$$\begin{aligned}&(rd+1)\int_0^\infty C\frac{z^{rd+1}}{rd+1}\exp\left(-C\frac{z^{rd+1}}{rd+1}\right)dz\\ &= (rd+1)\int_0^\infty ue^{-u}dz\\ &= (rd+1)\int_0^\infty ue^{-u}\left(\frac{u^{-rd/(rd+1)}}{rd+1}\left(\frac{rd+1}{C}\right)^{1/(rd+1)}du\right)\\ &= \left(\frac{rd+1}{C}\right)^{1/(rd+1)}\int_0^\infty u^{1/(rd+1)}e^{-u}du\end{aligned}$$

$$\begin{aligned}&= \left(\frac{rd+1}{C}\right)^{1/(rd+1)}\Gamma\left(\frac{rd+2}{rd+1}\right)\\ &= \left(\frac{rd+1}{C}\right)^{1/(rd+1)}\Gamma\left(\frac{rd+2}{rd+1}\right)\\ &\le \left(\frac{rd+1}{C}\right)^{1/(rd+1)},\end{aligned} \tag{42}$$

where (42) follows because $\Gamma(x) \le 1$ for $x \le 2$. ■

C. Proof of Corollary 2

To prove the "only if" direction, note that Theorem 3 asserts that there must be a constant $K > 0$ such that

$$r_n d_n \ge K\log\binom{n}{r_n} \ge Kr_n\log\frac{n}{r_n}.$$

Thus, by the assumption that $r_n = O(n^{1-\epsilon})$ for some $\epsilon > 0$,

$$d_n \ge K\log\frac{n}{r_n} = \Omega(\log n).$$

To prove the "if" direction, suppose there are positive constants K, ϵ, c such that for n sufficiently large

$$\begin{aligned}d_n &\ge K\log n\\ r_n &\le cn^{1-\epsilon}.\end{aligned}$$

Since reducing d_n can only adversely affect performance, we can assume without loss of generality that $d_n = K\log n$, and so for n sufficiently large, there exists some constant C such that

$$\begin{aligned}&(r_nd_n+1)\left((r_n)^{r_n}\binom{n}{r_n}\right)^{1/(r_nd_n+1)}\\ &\quad\le (r_nd_n+1)\left((r_n)^{r_n}\left(\frac{en}{r_n}\right)^{r_n}\right)^{1/(r_nd_n+1)}\\ &\quad\le Cr_nd_nn^{r_n/(r_nd_n+1)}\\ &\quad\le Cr_nd_nn^{1/d_n}\\ &\quad= Cr_n(K\log n)n^{1/(K\log n)}\\ &\quad\le C2^{1/K}\frac{K}{\epsilon}r_n\log\frac{n}{r_n}\\ &\quad\le C2^{1/K}\frac{K}{\epsilon}c\log\binom{n}{r_n}.\end{aligned}$$

An application of Theorem 2 completes the proof.

D. Details for Proof of Theorem 5

In this section, we provide the proof of Lemma 5, which has been restated below for convenience.

Lemma 5. Let p be a valid distribution corresponding to an α-optimal (r, d, n)-locally decodable code. Then there exist constants $\epsilon(\alpha), \delta(\alpha)$ independent of r, d, n, and $S \subseteq \mathbb{N}$ such that

1) $\alpha^{rd}\frac{p(k)}{p_{\max}(k)} \ge \epsilon(\alpha)\ \forall\ k \in S$, and
2) $\sum_{k\in S}p(k) \ge \delta(\alpha)$.

Proof: Since $\mathbb{E}_p[\ell(\mathsf{c}(X^n))] \le \alpha\left(\frac{rd+1}{4e}\right)\left(\binom{n}{r}^{1/(rd+1)}-1\right)$, an application of Markov's inequality for some fixed constant c yields that $\Pr_p[\ell(\mathsf{c}(X^n)) \le (\alpha+c)\left(\frac{rd+1}{4e}\right)\left(\binom{n}{r}^{1/(rd+1)}-1\right)] \ge \frac{c}{\alpha+c}$. For ease of notation, we define

$K_{\alpha,c} \triangleq (\alpha+c)\left(\frac{rd+1}{4e}\right)\left(\binom{n}{r}^{1/(rd+1)} - 1\right)$]. We use the notation $\alpha^{rd}p(k) \ll p_{\mathsf{max}}(k)$ to mean $\alpha^{rd}\frac{p(k)}{p_{\mathsf{max}}(k)} \to 0$ as any of $r, d, n \to \infty$.

We prove Lemma 5 by contradiction. Assume that it is false, i.e., for all sets $S \subseteq [K_{\alpha,c}]$ such that $\sum_{k\in S} p(k) \geq \delta$, for some constant δ independent of r, d, n, there exists some $k \in S$ such that $\alpha^{rd}p(k) \ll p_{\mathsf{max}}(k)$. Note that at least one set such that $\sum_{k\in S} p(k) \geq \delta$ must exist, since $\sum_{k\in[K_{\alpha,c}]} p(k) \geq \frac{c}{\alpha+c}$. Now consider one such set S, and remove from it all elements k such that $\alpha^{rd}p(k) \ll p_{\mathsf{max}}(k)$ to form the set S'. Since S' cannot be such that $\sum_{k\in S'} p(k) \geq \delta'$ for any fixed $\delta' > 0$, we must have that $\sum_{k\in S\setminus S'} p(k) \geq \delta$. In other words, there exists some set $S \subseteq [K_{\alpha,c}]$ such that $\alpha^{rd}p(k) \ll p_{\mathsf{max}}(k)$ $\forall$ $k \in S$, and $\sum_{k\in S} p(k) \geq \delta$. We complete the proof by arguing that no such set can exist.

Recall that $\sum_{k\in S} p(k) \geq \delta$. Along with the constraint that $\alpha^{rd}p(k) \ll p_{\mathsf{max}}(k)$, this implies that

$$\sum_{k\in S} \frac{p_{\mathsf{max}}(k)}{\alpha^{rd} f_k(n,r,d)} \geq \delta, \tag{43}$$

for some sequence of functions $\{f_k(n,r,d)\}_{k\in S}$ such that $f_k \to \infty$ as one of $r, d, n \to \infty$ $\forall$ $k \in S$. Now defining $g(n,r,d) \triangleq \min_k f_k(n,r,d)$ as the pointwise minimum of all the functions, we have:

$$\begin{aligned}
&\sum_{k\in S} \frac{p_{\mathsf{max}}(k)}{\alpha^{rd} f_k(n,r,d)}\\
&\quad\leq \frac{1}{\alpha^{rd} g(n,r,d)} \sum_{k\in S} p_{\mathsf{max}}(k)\\
&\quad\leq \frac{1}{\alpha^{rd} g(n,r,d)} \sum_{k\in[K_{\alpha,c}]} p_{\mathsf{max}}(k)\\
&\quad= \frac{1}{\alpha^{rd} g(n,r,d)} \sum_{k\in[K_{\alpha,c}]} \frac{\max_{v\leq rd}\binom{2k}{v}}{\binom{n}{r}}\\
&\quad\leq \frac{1}{\alpha^{rd} g(n,r,d)\binom{n}{r}} (rd+1)^{1/2} 2^{rd} \frac{\left(K_{\alpha,c}+2+\frac{rd+1}{2e}\right)^{rd+1}}{(rd+1)!},
\end{aligned} \tag{44}$$

where (44) follows from (12) by substituting $M \leftarrow K_{\alpha,c}$. Recalling the definition of $K_{\alpha,c}$, we now have

$$\sum_{k\in S} \frac{p_{\mathsf{max}}(k)}{\alpha^{rd} f_k(n,r,d)} \leq \frac{\alpha}{g(n,r,d)}.$$

Notice however that this is a contradiction to (43), since $g(n,r,d) \to \infty$. Lemma 5 is hence proved. ∎

References

[1] A. Pananjady and T. Courtade, "Compressing sparse sequences under local decodability constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2979–2983.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.

[3] D. A. Huffman *et al.*, "A method for the construction of minimum redundancy codes," *Proc. IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.

[4] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 3, pp. 337–343, May 1977.

[5] D. S. Pavlichin, T. Weissman, and G. Yona, "The human genome contracts again," *Bioinformatics*, vol. 29, no. 17, pp. 2199–2202, 2013.

[6] S. Deorowicz, A. Danek, and S. Grabowski, "Genome compression: A novel approach for large collections," *Bioinformatics*, vol. 29, no. 20, pp. 2572–2578, 2013.

[7] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[8] P. Elias and R. A. Flower, "The complexity of some simple retrieval problems," *J. ACM*, vol. 22, no. 3, pp. 367–379, 1975.

[9] P. B. Miltersen, "The bit probe complexity measure revisited," in *Proc. Annu. Symp. Theor. Aspects Comput. Sci.*, 1993, pp. 662–671.

[10] P. B. Miltersen, "Cell probe complexity—A survey," in *Proc. 19th Conf. Found. Softw. Technol. Theor. Comput. Sci. (FSTTCS)*, 1999, pp. 1–25.

[11] A. C.-C. Yao, "Should tables be sorted?" *J. ACM*, vol. 28, no. 3, pp. 615–628, 1981.

[12] H. Buhrman, P. B. Miltersen, J. Radhakrishnan, and S. Venkatesh, "Are bitvectors optimal?" *SIAM J. Comput.*, vol. 31, no. 6, pp. 1723–1744, 2002.

[13] P. K. Nicholson, V. Raman, and S. S. Rao, "A survey of data structures in the bitprobe model," in *Space-Efficient Data Structures, Streams, and Algorithms*. Berlin, Germany: Springer, 2013, pp. 303–318.

[14] N. Alon and U. Feige, "On the power of two, three and four probes," in *Proc. 20th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2009, pp. 346–354.

[15] J. Radhakrishnan, V. Raman, and S. S. Rao, "Explicit deterministic constructions for membership in the bitprobe model," in *Proc. Eur. Symp. Algorithms*, 2001, pp. 290–299.

[16] E. Viola, "Bit-probe lower bounds for succinct data structures," *SIAM J. Comput.*, vol. 41, no. 6, pp. 1593–1604, 2012.

[17] M. Lewenstein, J. I. Munro, P. K. Nicholson, and V. Raman, "Improved explicit data structures in the bitprobe model," in *Proc. Eur. Symp. Algorithms*, 2014, pp. 630–641.

[18] M. Garg and J. Radhakrishnan, "Set membership with a few bit probes," in *Proc. 26th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2015, pp. 776–784.

[19] A. Mazumdar, V. Chandar, and G. Wornell, "Local recovery in data compression for general sources," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2015, pp. 2984–2988.

[20] V. Chandar, D. Shah, and G. W. Wornell, "A locally encodable and decodable compressed data structure," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2009, pp. 613–619.

[21] A. Makhdoumi, S.-L. Huang, Y. Polyanskiy, and M. Medard. (Aug. 2013). "On locally decodable source coding." [Online]. Available: https://arxiv.org/abs/1308.5239

[22] H. Zhou, D. Wang, and G. Wornell, "A simple class of efficient compression schemes supporting local access and editing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2014, pp. 2489–2493.

[23] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1670–1672, Sep. 1994.

[24] J. Katz and L. Trevisan, "On the efficiency of local decoding procedures for error-correcting codes," in *Proc. 32nd Annu. ACM Symp. Theory Comput.*, 2000, pp. 80–86.

[25] K. Yamamoto, "Logarithmic order of free distributive lattice," *J. Math. Soc. Jpn.*, vol. 6, nos. 3–4, pp. 343–353, 1954.

[26] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson, "On data structures and asymmetric communication complexity," in *Proc. 27th Annu. ACM Symp. Theory Comput.*, 1995, pp. 103–111.

Ashwin Pananjady is a PhD candidate in the Department of Electrical Engineering and Computer Sciences (EECS) at the University of California, Berkeley. He received the B.Tech. degree (with honors) in electrical engineering from IIT Madras in 2014. His research interests include statistics, optimization, and information theory. He is a recipient of the Governor's Gold Medal from IIT Madras.

Thomas A. Courtade received the B.Sc. degree (summa cum laude) in electrical engineering from Michigan Technological University, Houghton, MI, USA, in 2007, and the M.S. and Ph.D. degrees from the University of California, Los Angeles (UCLA), CA, USA, in 2008 and 2012, respectively. He is an Assistant Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. Prior to joining UC Berkeley in 2014, he was a Postdoctoral Fellow supported by the NSF Center for Science of Information. Prof. Courtade's honors include a Distinguished Ph.D. Dissertation Award and an Excellence in Teaching Award from the UCLA Department of Electrical Engineering, and a Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory. He was the recipient of a Hellman Fellowship in 2016.