

An Upper Bound On the Size of Locally Recoverable Codes

Viveck Cadambe *Member, IEEE* and Arya Mazumdar *Member, IEEE*

Abstract

In a *locally recoverable* or *repairable* code, any symbol of a codeword can be recovered by reading only a small (constant) number of other symbols. The notion of local recoverability is important in the area of distributed storage where a most frequent error-event is a single storage node failure (erasure). A common objective is to repair the node by downloading data from as few other storage node as possible. In this paper, we bound the minimum distance of a code in terms of its length, size and locality. Unlike previous bounds, our bound follows from a significantly simple analysis and depends on the size of the alphabet being used. It turns out that the binary Simplex codes satisfy our bound with equality; hence the Simplex codes are the first example of a optimal binary locally repairable code family. We also provide achievability results based on random coding and concatenated codes that are numerically verified to be close to our bounds.

I. INTRODUCTION

The increased demand of cloud computing and storage services in current times has led to a corresponding surge in the study and deployment of erasure-correcting codes, or simply erasure codes, for distributed storage systems. In the information and coding theory community, this

A preliminary version of this work has appeared in IEEE International Symposium on Network Coding (NetCod), 2013. This work was supported in part by NSF CCF 1318093.

A. Mazumdar is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 (Email: arya@umn.edu).

Viveck Cadambe is with the Department of Electrical Engineering, Pennsylvania State University, University Park PA 16802 (Email: viveck@engr.psu.edu).

has led to the research of some new aspects of codes particularly tailored to the application to storage systems. The topic of interest of this paper is the locality of repair of erasure codes.

It is well known that an erasure code with length n , dimension k and minimum distance d , or an (n, k, d) code, can recover from *any* set of $d - 1$ erasures. In addition, the code is said to have *locality* r if any *single* erasure can be recovered from some set of r symbols of the codeword. From an engineering perspective, when an (n, k, d) code is used to store information in n storage nodes, the parameter d represents the worst-case (node) failure scenario from which the storage system can recover. The parameter r , on the other hand, represents the efficiency of recovery from a (relatively) more commonly occurring scenario - a single node failure. It is therefore desirable to have a large value of d and a small value of r . Much literature in classical coding theory has been devoted to understanding the largest possible value of d - the minimum distance - when the parameters (n, k) are fixed; the well-known results from this body [23] of work include the Singleton bound, and code constructions that achieve this bound (such as Reed-Solomon codes). The study of minimizing the locality, r , was initiated recently in [8], [17] and furthered in [11], [12], [18], [19], [21], [24]. The key discovery of [8], [18], [20] is that, for any (n, k, d) code with locality r , the following bound is satisfied:

$$d \leq n - k - \lceil k/r \rceil + 2. \quad (1)$$

The above bound is a generalization of the Singleton bound to include the locality of the codeword, r ; when $r = k$, the above bound collapses to the classical Singleton bound. In addition, through an invocation of the *multicast* capacity of wireline networks via random network coding, reference [18] showed that the bound (1) is indeed tight for a sufficiently large field size. Intuitively speaking, the bound (1) implies that there is a *cost* to locality; the smaller the locality, r , the smaller the minimum-distance d . Code constructions that achieve the above bound based on Reed-Solomon codes, among other techniques, have been recently discovered in [10], [12],

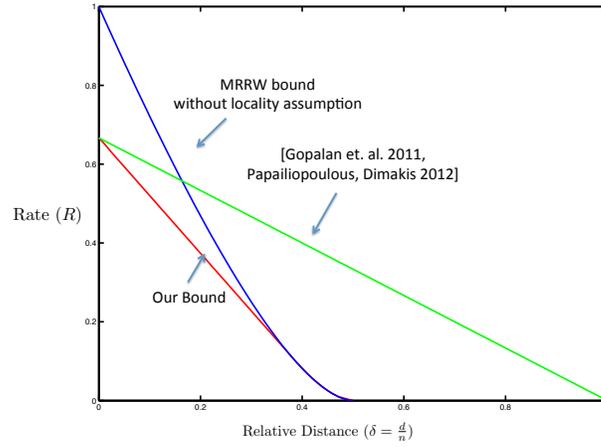


Fig. 1. A depiction of our bound through the trade-off between the *rate*, k/n and *relative distance*, d/n , for binary codes ($q=2$) for large values of n , with locality $r = 2$. The curves plotted are upper bounds on the achievable rates; the plot clearly demonstrates that our upper bound, that uses MRRW bound as a black-box, is better than the previously known bounds on the rate, for a given relative distance.

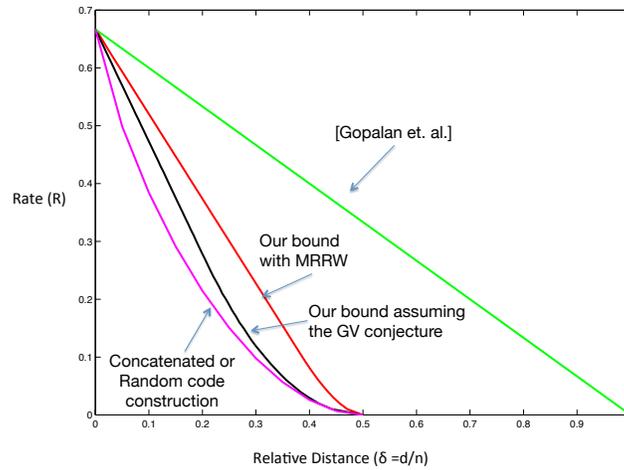


Fig. 2. A depiction of our achievability result (6) through the trade-off between the *rate*, k/n and *relative distance*, d/n , for binary codes ($q=2$) for large values of n , with locality $r = 2$. We compare this achievable rate with our upper bounds: assuming respectively MRRW bound and the GV bound as the optimal rate for error-correcting codes. If the GV bound were true rate-distance trade-off, then our achievability scheme is very good for large distances.

[13], [18], [21], [27]¹. Missing from these works is a formal study of the impact of an important parameter - the size of the alphabet of the code. Codes over small alphabets are the central subject of classical coding theory, and are of interest in the application of storage because of their implementation ease. In this paper, we remove the restriction of the large alphabet size from the study of locality of codeword symbols. In particular, we study the impact of the alphabet size on the locality of a code that has a fixed rate.

A. *Our Contribution:*

The main contribution of this paper is an upper bound on the minimum distance on the code with a fixed locality that is dependent on the size of the alphabet. While the technical statement of our bound is discussed later (in Theorem 1), it is worth noting here that our bound, which is applicable for any feasible alphabet-size and any feasible (n, k) , is least as good as the bound of [8], [18] for all parameters. Recall that even in the absence of locality constraints, finding the largest possible minimum distance of a code with a fixed rate over a fixed-size alphabet remains an open problem in general. Our main result uses this quantity - the largest possible minimum distance of an (n, k) code over a given alphabet size - albeit unknown, as a parameter to obtain a bound under locality constraints. As a consequence, our bound is more stringent than the classical (locality-unaware) bounds such as the McIiece-Rodemich-Rumsey-Welch (MRRW) bounds since they form a special case of our bound (unrestricted locality). The role of the alphabet size on the rate of the code is highlighted in the plot of Fig. 1, where we compare our bounds with existing bounds. Since certain code constructions in previous works are based on multicast codes over networks [18], our result can be interpreted as the demonstration of the impact of alphabet size on the rates of multicast *network codes* for certain networks. Finally, we discuss some achievable constructions in this paper and compare them with our bound. In particular, in Sec. V, we will show that the family of Simplex codes (dual of Hamming codes) are optimal locally recoverable

¹Recent literature has extended the study of locally recoverable codes to include security constraints [21], notions of repair bandwidth [12], [13], multiple local repair alternatives [20], and probabilistic erasure models [16] (See [5] for a survey). In this work, we concern ourselves with the original notion of locality as described in [8], [18].

codes since they meet our new bound. Constructions based on random coding and concatenated codes are also provided that achieve a rate-distance tradeoff asymptotically close to the optimal possible (Section VI). After the initial publication of our work [4], interesting generalizations of the Simplex code construction of Sec. V have been studied in [9], [25], [28]. Also worth noting is the elegant construction of an optimal family of codes by Tamo and Barg [26]. Nonetheless, the search for optimal locally recoverable codes remains open even for the binary alphabet. Our randomized construction of Section VI outperform previous code constructions for smaller alphabet sizes including the binary alphabet. We end with a discussion on construction based on LDPC-type codes and list decoding.

Notation: Sets are denoted by calligraphic letters and vectors are denoted by bold font. Consider an element $\mathbf{X} \in \mathcal{A}^n$, where \mathcal{A} is an arbitrary finite set. The notation $X_i \in \mathcal{A}$ denotes the i th co-ordinate of the tuple \mathbf{X} . For any set $\mathcal{R} \subseteq \{1, 2, \dots, n\}$, the notation $\mathbf{X}_{\mathcal{R}} \in \mathcal{A}^{|\mathcal{R}|}$ denotes the projection of $\mathbf{X} \in \mathcal{A}^n$ on to the co-ordinates corresponding to \mathcal{R} . For $\mathbf{X}, \mathbf{Y} \in \mathcal{A}^n$, the Hamming distance $\Delta_H(\mathbf{X}, \mathbf{Y})$ is the cardinality of the set $\{m : X_m \neq Y_m\}$.

II. SYSTEM MODEL: LOCALLY RECOVERABLE CODES

A code \mathcal{C} with length n over alphabet \mathcal{Q} consists of $|\mathcal{C}|$ codewords denoted as

$$\mathcal{C} = \{\mathbf{X}^n(1), \mathbf{X}^n(2), \dots, \mathbf{X}^n(|\mathcal{C}|)\},$$

where $\mathbf{X}^n(i) \in \mathcal{Q}^n, \forall i$. The *dimension of the code*, denoted by k is defined as $k \triangleq \frac{\log |\mathcal{C}|}{\log |\mathcal{Q}|}$, and the rate of the code denoted as R is defined as $R \triangleq \frac{k}{n}$. An (n, k, d) -code over \mathcal{Q} is an n length code \mathcal{C} with dimension k such that the *minimum distance* is d , i.e., with

$$d = \min_{\mathbf{X}^n, \mathbf{Y}^n \in \mathcal{C}, \mathbf{X}^n \neq \mathbf{Y}^n} \Delta_H(\mathbf{X}^n, \mathbf{Y}^n).$$

We refer to $\delta \triangleq \frac{d}{n}$ as the *relative distance* of the code.

Definition 1: An (n, k, d) -code is said to be *r-locally recoverable* if for every i such that $1 \leq i \leq n$, there exists a set $\mathcal{R}_i \subset \{1, 2, \dots, n\} \setminus \{i\}$ with $|\mathcal{R}_i| = r$ such that for any two codewords \mathbf{X}, \mathbf{Y} satisfying $X_i \neq Y_i$, we have $\mathbf{X}_{\mathcal{R}_i} \neq \mathbf{Y}_{\mathcal{R}_i}$.

Informally speaking, this means that an erasure of the i th coordinate of the codeword can be recovered by accessing the coordinates associated with \mathcal{R}_i . Hence any erased symbol can be recovered by probing at most r other coordinates.

III. BOUND ON MINIMUM DISTANCE FOR LOCAL RECOVERY

Given parameters n, d, q , let

$$k_{\text{opt}}^{(q)}(n, d) = \max \frac{\log |\mathcal{C}|}{\log q},$$

where the maximization is over all possible n -length codebooks \mathcal{C} with minimum distance d , over some alphabet \mathcal{Q} where $|\mathcal{Q}| = q$. Informally speaking, $k_{\text{opt}}^{(q)}(n, d)$ is the largest possible dimension of an n -length code, for a given alphabet size q and a given minimum distance d . The determination of $k_{\text{opt}}^{(q)}$ is a classical open problem in coding theory. We also know that k_{opt} satisfies the Singleton bound:

$$k_{\text{opt}}^{(q)}(n, d) \leq n - d + 1, \forall q \in \mathbb{Z}_+.$$

References [8], [18], generalized the above bound under locality constraints as Def. 1. However, it is well known that the Singleton bound is not tight in general, especially for small values of q . The goal of this paper is to derive a bound on the dimension of an r -locally recoverable code in terms of $k_{\text{opt}}^{(q)}$. Our main result is the following.

Theorem 1: For any (n, k, d) code over \mathcal{Q} that is r -locally recoverable

$$k \leq \min_{t \in \mathbb{Z}_+} \left[tr + k_{\text{opt}}^{(q)}(n - t(r + 1), d) \right], \quad (2)$$

where $q = |\mathcal{Q}|$.

Our bound applies to general (including non-linear) codes, as opposed to only linear codes. Note that, the minimizing value of t in (2), t^* , must satisfy,

$$t^* \leq \min \left\{ \left\lceil \frac{n}{r+1} \right\rceil, \left\lceil \frac{k}{r} \right\rceil \right\}.$$

This is true because, 1) for $t \geq \left\lceil \frac{n}{r+1} \right\rceil$, the objective function of the optimization of (2) becomes linearly growing with t ; 2) for $t \geq \left\lceil \frac{k}{r} \right\rceil$, the right hand side of (2) is greater than k .

The bound of [8], [18], i.e. (1), is weaker than the bound of Theorem 1. To prove this claim, let us show that, if a (n, k, d, r) -tuple does not satisfy (1), then it will not satisfy (2).

If possible, let the tuple (n, k, d, r) violate (1), i.e., let

$$d > n - k - \lceil k/r \rceil + 2.$$

This sets the following chain of implications.

$$\begin{aligned} & \min_{t \in \mathbb{Z}_+} \left[tr + k_{\text{opt}}^{(q)}(n - t(r + 1), d) \right] \\ & \leq \lfloor (k - 1)/r \rfloor r + \max\{n - \lfloor (k - 1)/r \rfloor (r + 1) - d + 1, 0\} \\ & = \max\{n - \lfloor (k - 1)/r \rfloor - d + 1, \lfloor (k - 1)/r \rfloor r\} \\ & < \max\{n - \lfloor (k - 1)/r \rfloor - n + k + \lceil k/r \rceil - 2 + 1, k\} \\ & = \max\{-\lfloor (k - 1)/r \rfloor + k + \lceil k/r \rceil - 1, k\} \\ & = k, \end{aligned}$$

which means (2) is not satisfied by this tuple as well.

Notice that the above chain of implications came from plugging in the Singleton bound on $k_{\text{opt}}^{(q)}$. We shall apply bounds that are dependent on q and stronger than the Singleton bound on $k_{\text{opt}}^{(q)}$ to effectively obtain tighter bounds on (1) later in this paper. We shall first present an overview of the proof of Theorem 1. For purposes of the proof, for a given n length code \mathcal{C} we define the function $H(\cdot)$ as follows

$$H(\mathcal{I}) = \frac{\log |\{\mathbf{X}_{\mathcal{I}} : \mathbf{X} \in \mathcal{C}\}|}{\log |\mathcal{Q}|},$$

for any set $\mathcal{I} \subseteq \{1, 2, \dots, n\}$.

Remark 1: In the language used in [18], $H(\mathcal{I})$ would denote the ‘‘entropy’’ associated with $\mathbf{X}_{\mathcal{I}}$. Here, the above definition is appropriate since our modeling is adversarial, i.e., we do not presuppose any distribution on the messages or the codebook (see, [16] where such assumptions have been made). However, the behavior of the function $H(\cdot)$ is similar to the entropy function; for instance it satisfies submodularity, i.e., $H(\mathcal{I}_1) + H(\mathcal{I}_2) \geq H(\mathcal{I}_1 \cup \mathcal{I}_2) + H(\mathcal{I}_1 \cap \mathcal{I}_2)$

Theorem 1 follows from Lemma 1 and Lemma 2 stated next.

Lemma 1: Consider an (n, k, d) -code over alphabet \mathcal{Q} that is r -locally recoverable. Then, $\forall 1 \leq t \leq k/r, t \in \mathbb{Z}$ there exists a set $\mathcal{I} \subseteq \{1, 2, \dots, n\}, |\mathcal{I}| = t(r + 1)$ such that $H(\mathcal{I}) \leq tr$.

Lemma 2: Consider an (n, k, d) -code over \mathcal{Q} where there exists a set $\mathcal{I} \in \{1, 2, \dots, n\}$ such that $H(\mathcal{I}) \leq m$. Then there exists a $(n - |\mathcal{I}|, (k - m)^+, d)$ code over \mathcal{Q} .

The above lemmas are proved in the appendix.

IV. APPLICATIONS OF THM. 1

In this section, we apply classical bounds for k_{opt} to Theorem 1. To enable a clean analysis, we look at the regime where $n \rightarrow \infty$. In particular we set $R = k/n, \delta = d/n$ and obtain bounds on the trade-off between (R, δ) as r is fixed and $n \rightarrow \infty$. We first apply the Plotkin bound on k_{opt} and obtain an analytical characterization of the (R, δ) trade-off with dependence on the alphabet-size, q ; in particular, we demonstrate a *distance-expansion* penalty as a result of the limit on alphabet size. To obtain a tighter locality-aware bound, we then use the MRRW bound for k_{opt} to numerically obtain the plot of Fig. 1.

To begin, observe that dividing the Singleton bound n and letting $n \rightarrow \infty$, it can be written as

$$R \leq 1 - \delta + o(1)$$

Similarly, the bound of [8], [18] can be written as:

$$\begin{aligned} \delta &\leq 1 - \frac{rR}{r+1} + o(1). \\ \Rightarrow R &\leq \frac{r}{r+1}(1 - \delta) + o(1) \end{aligned} \tag{3}$$

The plot of the above bound is placed in Fig. 1 for $r = 2$. The *cost* of the locality limit above therefore is the factor of $r/(r + 1)$ over the Singleton bound. We are now ready to analyze the Plotkin Bound, adapted to Theorem 1.

Application of Plotkin Bound - Distance Expansion Penalty

Let us choose $t = \frac{1}{r+1}(n - \frac{d}{1-1/q})$ in Theorem 1. We have, for any (n, k, d) -code that is r -locally recoverable,

$$k \leq \frac{r}{r+1} \left(n - \frac{d}{1-1/q} \right) + k_{\text{opt}}^{(q)} \left(\frac{d}{1-1/q}, d \right)$$

It is known, from the Plotkin bound, $k_{\text{opt}}^{(q)} \left(\frac{d}{1-1/q}, d \right) \leq \log_q \frac{2qd}{1-1/q}$. See, for example, Sec. 2§2 of MacWilliams and Sloane [15], for a proof of this result for $q = 2$, which can be easily extended for larger alphabets. Hence,

$$k \leq \frac{r}{r+1} \left(n - \frac{d}{1-1/q} \right) + \log_q \frac{2qd}{1-1/q}. \quad (4)$$

Generally, this bound is better than (1). Notice that dividing the above by n and taking $n \rightarrow \infty$, we have

$$R = \frac{k}{n} \leq \frac{r}{r+1} \left(1 - \frac{\delta}{1-1/q} \right) + o(1),$$

whereas, observing the above, it can be noted that the effect of restricting q leads to a distance-expansion penalty of $\frac{1}{1-1/q}$, since the above bound is tantamount to shooting for a distance of $\delta/(1-1/q)$ w.r.t. (3).

Beyond the Plotkin bound

Recall that the MRRW bound is the tightest known bound for the rate-distance tradeoff in absence of locality constraints. We briefly describe an application of this bound for Theorem 1, i.e., when the locality is restricted to be equal to a number r ; it is this bound that is plotted in Fig. 1. We restrict our attention to binary codes ($q = 2$) and therefore the dependence on q is dropped in the notation.

Define $R_{\text{opt}}(\delta) \triangleq \lim_{n \rightarrow \infty} \frac{k_{\text{opt}}(n, \delta n)}{n}$. Dividing the bound of Theorem 1 by n we can get, as $n \rightarrow \infty$,

$$R \leq \min_{0 \leq x \leq r/(r+1)} x + (1 - x(1 + 1/r)) R_{\text{opt}} \left(\frac{\delta}{1 - x(1 + 1/r)} \right) \quad (5)$$

where $x = tr/n$. It is instructive to observe that, setting $x = 0$ above yields classical (locality-unaware) bounds. Setting $x = R$ above and writing out the Singleton bound for R_{opt} yields the bound of (3). Therefore the above bound is superior to all the classical (locality-unaware) bounds on $R(\delta)$ and the bound of (3) since these are special cases. Using the MRRW bound $R(y) \leq H_2(0.5 - \sqrt{y(1-y)}) + o(1)$ (where $H_q(x) = x \log_q(q-1) - x \log_q x - (1-x) \log_q(1-x)$ represents the q -ary entropy function), and numerically solving the optimization problem above (in a brute-force manner) yields our bounds for the rate-distance trade-offs for any given r . Deriving analytical insights for the optimization problem by application of bounds beyond the Plotkin bound is an area of future work.

Remark 2: While the MRRW bound is the best known upper bound on the rate given a relative distance, for binary codes, the best known achievable scheme, asymptotically as $n \rightarrow \infty$, is given by the Gilbert-Varshamov (GV) Bound. Indeed, it is a folklore conjecture in coding theory that the GV bound is the best achievable rate for binary code, asymptotically as the blocklength tends to infinity. Therefore, to evaluate the merit of binary locally recoverable code constructions, the use of the GV bound for the function R_{opt} in (5) has operational meaning.

V. SIMPLEX CODES AND TIGHTNESS OF THM. 1

For alphabet size exponential in the blocklength the bound of (1) has been shown to be achievable in [21], [27] by constructing explicit codes. Furthermore, recently, [26] has shown that an alphabet that is linear in the blocklength suffices to achieve the bound of 1. Hence (2) is tight for large alphabets. We will show that this bound is also achievable for small, in particular binary, alphabets by giving an example of explicit family of codes where (2) is met with equality. The family of codes is $[2^m - 1, m, 2^{m-1}]$ Simplex code, $m \in \mathbb{Z}_+$.

First, we derive, according to Thm. 1, the best possible locality a code with the parameters of Simplex code can have. Here, $n \equiv 2^m - 1$, $k \equiv m = \log_2(n + 1)$, $d \equiv 2^{m-1} = \frac{n+1}{2}$. We use $t = 2$, which satisfies

$$t \leq \min \left\{ \left\lceil \frac{n}{r+1} \right\rceil, \left\lceil \frac{k}{r} \right\rceil \right\},$$

as will be clear next. With this value and using Plotkin bound [15, Sec. 2§2]:

$$\begin{aligned} k = m &\leq 2r + k_{\text{opt}}(2^m - 1 - 2(r + 1), 2^{m-1}) \\ &\leq 2r + \log_2 \frac{2 \cdot 2^{m-1}}{2^m - 2^m + 1 + 2(r + 1)} \\ &= 2r + m - \log_2(2r + 3). \end{aligned}$$

That is,

$$2r \geq \log_2(2r + 3) \Rightarrow r \geq 2.$$

Hence according to Thm. 1, the best possible locality with the parameters of Simplex code is 2.

Next, we show that the Simplex code indeed has locality 2. This is shown by constructing a parity-check matrix of Simplex code, with every row having exactly 3 ones. Recall, the dual code of Simplex code is a $[2^m - 1, 2^m - 1 - m, 3]$ -Hamming code. We give a generator matrix of Hamming code that has only 3 ones per row. Let us index the columns of the generator matrix by $1, 2, \dots, 2^m - 1$, and use the notation (i, j, k) to denote the vector with exactly three 1's, located at positions i, j , and k . Then, the Hamming code has a generator matrix given by the row vectors $(i, 2^j, i + 2^j)$ for $1 \leq j \leq m - 1, 1 \leq i < 2^j$.

This gives the first example of a family of algebraic codes that are optimal in terms of local repairability.

VI. ACHIEVABILITY BOUNDS AND CONSTRUCTIONS

So far in this paper, we have provided upper bounds on on the rate achievable for a fixed locality, distance, and alphabet size. Constructions of locally recoverable codes is an interesting open question especially relevant to practice. To understand the related issues (briefly), consider the special case of binary codes ($q = 2$) where, in absence of locality constraints, the best known simple achievable scheme comes via the Gilbert-Varshamov (GV) bound: $R \geq 1 - H_2(\delta)$. Note that, to achieve a locality of r with a linear code, it is sufficient for the parity check matrix of the code to have the following property: for every column, there exists a row vector in the parity check matrix with a non-zero entry in that column, and a hamming weight that is no bigger than

$r + 1$. A simple construction for locally recoverable codes is constructed by taking the parity check matrix of a code that achieves the GV bound and add $\lceil \frac{n}{r+1} \rceil$ rows to it; each new row has $r + 1$ nonzero values and the support of all the (new) rows are disjoint. Clearly, this code has a locality of r . Note that this new code has rate:

$$R \geq 1 - H_2(\delta) - \frac{1}{r+1} = \frac{r}{r+1} - H_2(\delta).$$

For $\delta = 0$, the above clearly meets the outer bound of (1). However, the above achievable scheme does not meet our bound for larger values of δ . For example, in the regime of Fig. 1, i.e., $r = 2$, the above bound implies that $R = 0$ for $\delta \geq H_2^{-1}(2/3) \approx 0.18$. Clearly, this is not tight with our bound, where $R > 0$ as long as $\delta < 0.5$. This motivates the following question: what is largest possible (relative) distance of a code with non-zero rate, for a fixed locality and alphabet size? We answer this question in the next section. In particular, we provide two families code constructions that perform well from the perspective of our bounds.

A. Random codes

Suppose $r + 1$ divides n . Construct a random code of length n in the following way. Let $X_{i,j}, 1 \leq i \leq \frac{n}{r+1}, 1 \leq j \leq r$, are randomly and uniformly chosen from \mathbb{F}_q . Let $X_{i,r+1} = \sum_{j=1}^r X_{i,j}$, where the addition is over \mathbb{F}_q .

Assume, $X_{i,j}, 1 \leq i \leq \frac{n}{r+1}, 1 \leq j \leq r + 1$, is a codeword of a random code. We choose such a random code consisting of M independent codewords $\mathbf{X}_1, \dots, \mathbf{X}_M$. The length, locality and dimension of any code in this random ensemble is n, r and $k = \log_q M$ respectively.

Theorem 2: There exists codes in the above ensemble with minimum distance at least d , where d is given by

$$\frac{k}{n} = 1 - \max_{0 \leq x \leq 1} \left[\log_q(1 + x(q-1)) + \frac{1}{r+1} \log_q \left(1 + (q-1) \left(\frac{1-x}{1+x(q-1)} \right)^{r+1} \right) - \frac{d}{n} \log_q x \right]. \quad (6)$$

The proof of this theorem follows the usual random coding methods and the calculation is quite similar to that of the following Theorem 3, that proclaims the same result for a linear code ensemble. We delegate the proof to the appendix.

B. Concatenated codes

One approach to construct a locally recoverable code over alphabet size q much smaller than blocklength n is to use Forney's concatenated codes [6].

Consider a concatenated code with an outer extended Reed-Solomon code over alphabet \mathbb{F}_{q^r} , length $n_o = q^r$, and dimension k_o . The minimum distance of the code is $d_o = q^r - k_o + 1$. The inner code is a simple q -ary parity check code of length $r + 1$ (i.e., dimension r). The overall code has length $n = (r + 1)q^r$, dimension $k = k_o r$ and distance

$$\begin{aligned} d &= 2(q^r - k_o + 1) \\ &= 2\left(\frac{n}{r+1} - \frac{k}{r} + 1\right) \\ \Rightarrow \frac{k}{n} &= \frac{r}{r+1} - \frac{r}{n}\left(\frac{d}{2} - 1\right). \end{aligned}$$

Comparing with (4), we conclude that this construction has some merit for small values of r .

Using concatenated code with a *random* linear outer code, we show a much tighter achievability result. Indeed, the following is true.

Theorem 3: There exists an infinite family of $[n, k, d]_q$ concatenated codes with locality r , such that (6) is satisfied.

Proof: We use an outer random q^r -ary linear code of length $\frac{n}{r+1}$ and dimension $\frac{k}{r}$. The inner code is a q -ary single parity-check code of length $r + 1$. The overall q -ary code has length n , dimension k and locality r .

For the encoding procedure, any vector in $\mathbb{F}_q^k \setminus \{0\}$ is first mapped to a vector in $\mathbb{F}_{q^r}^{k/r} \setminus \{0\}$ and then encoded to a codeword of the outer code. In the next step, the symbols of the codeword (of the outer code) are mapped to codewords of the inner code.

Because the outer code is random linear, for any $\mathbf{u} \in \mathbb{F}_q^k \setminus \{0\}$, the corresponding q -ary codeword is going to have Hamming weight W , with,

$$W = X_1 + X_2 + \cdots + X_{\frac{n}{r+1}},$$

where $X_i \sim X$ are independent identical random variables such that

$$\Pr(X = j) = \begin{cases} \frac{1}{q^r} \binom{r+1}{j} \frac{q-1}{q} \left((q-1)^{j-1} + 1 \right) & \text{for even } j \\ \frac{1}{q^r} \binom{r+1}{j} \frac{q-1}{q} \left((q-1)^{j-1} - 1 \right) & \text{for odd } j. \end{cases} \quad (7)$$

We used the weight distribution of q -ary single parity-check code from [23, E.g. 4.6]. Similar reasoning has been followed in [1, Prop. 1] where more general inner codes were considered. It is instructive to note that when $q = 2$, the above equation implies that all even weight codewords are equiprobable, and the odd weight codewords have zero probability. Now, for any $t > 0$,

$$\begin{aligned} \mathbb{E}e^{-tX} &= \frac{q-1}{q^{r+1}} \sum_{j=0}^{r+1} e^{-tj} \binom{r+1}{j} \left((q-1)^{j-1} + (-1)^j \right) \\ &= \frac{1}{q^{r+1}} \left((1 + e^{-t}(q-1))^{r+1} + (q-1)(1 - e^{-t})^{r+1} \right). \end{aligned}$$

Evidently, for any $t > 0$,

$$\begin{aligned} \Pr(W < d) &= \Pr\left(\sum_{i=1}^{n/(r+1)} X_i < d \right) \\ &= \Pr\left(e^{-t \sum_{i=1}^{n/(r+1)} X_i} > e^{-td} \right) \\ &\leq e^{td} (\mathbb{E}e^{-tX})^{\frac{n}{r+1}}. \end{aligned}$$

Therefore, the average number of codewords of weight less than d is at most

$$\begin{aligned} \min_{0 \leq t} q^k e^{td} (\mathbb{E}e^{-tX})^{\frac{n}{r+1}} &= q^{k-n} \min_{0 \leq t} e^{td} \left((1 + e^{-t}(q-1))^{r+1} + (q-1)(1 - e^{-t})^{r+1} \right)^{n/(r+1)} \\ &= q^{k-n} \min_{0 \leq t} e^{td} (1 + e^{-t}(q-1))^n \left(1 + (q-1) \left(\frac{1 - e^{-t}}{1 + e^{-t}(q-1)} \right)^{r+1} \right)^{\frac{n}{r+1}} \\ &= q^{k-n} \min_{0 \leq x \leq 1} x^{-d} (1 + x(q-1))^n \left(1 + (q-1) \left(\frac{1 - x}{1 + x(q-1)} \right)^{r+1} \right)^{\frac{n}{r+1}}. \end{aligned}$$

As long as this number is less than 1, we must have a code in our ensemble that has minimum distance at least d . This proves the theorem. \blacksquare

It is not immediately apparent as to how tight the bound of (6) is. To see this, let us substitute $x = \frac{d}{(q-1)(n-d)}$, to have

$$\frac{k}{n} = 1 - H_q\left(\frac{d}{n}\right) - \frac{1}{r+1} \log_q \left(1 + (q-1) \left(1 - \frac{d}{n} \cdot \frac{q}{q-1} \right)^{r+1} \right). \quad (8)$$

It is clear that at $d = 0$, $\frac{k}{n} = \frac{r}{r+1}$ and at $d = n(1 - 1/q)$, $\frac{k}{n} = 0$. At least at these two points, thus, the bound of (6) exactly matches the upper bound of (2). We have plotted the bound of (6) by numerically optimizing over the parameter x in Figure 2. In Figure 2, we compare this achievability result with our upper bound, assuming the GV conjecture (that the GV bound is the asymptotically optimal achievable rate for a binary error-correcting code).

VII. DISCUSSIONS

A. LDPC codes

As *low density parity check matrix* (LDPC) codes are by definition locally repairable - any construction of LDPC codes provides locally recoverable codes. In particular, it is to be noted that there are a number of ensembles of LDPC codes, either based on random graphs or expander graphs, that have been extensively analyzed for their rate and distance trade-off [2]. It has been observed by Gallager [7] and others that the ensemble average distance of such codes approach the Gilbert-Varshamov bound as the degree of the parity-check graph grows. These codes also guarantee multiple recovering sets for each codeword symbol.

As an example, we can take the ensemble of hypergraph codes with Hamming codes as component codes from [2, Thm. 4] and the instances presented therein (see also, [3], [14]). Using a $[15, 11, 3]$ Hamming code as local code and a hypergraph with 3 parts, we are able to construct a code of rate 0.2, relative distance 0.2307 (the GV relative distance for this rate is 0.2430), local repairability $r = 11$ and three repair groups for each symbol. At the same time this codes support cooperative local repair [22], that is, there can be at most two erasures per repair group, that can still be locally corrected.

B. List decoding

Our upper-bounding techniques can be extended to bound the local recoverability of a list decodable code. Let $A_L^{(q)}(n, s)$ be the maximum possible size of a code \mathcal{C} such that for any ball of radius s in \mathbb{F}_q^n , there exist at most L codewords of \mathcal{C} . Such codes are called (s, L) -list decodable codes.

Theorem 4: Let \mathcal{C} be an q -ary (s, L) -list decodable code with length n , dimension k and local repairability r . Then

$$k \leq \min_{t \in \mathbb{Z}_+} \left[tr + k_L^{(q)}(n - t(r + 1), s) \right], \quad (9)$$

where $k_L^{(q)}(n, s) = \log_q A_L^{(q)}(n, s)$, and the minimizing value of t in (2), t^* , must satisfy,

$$t^* \leq \min \left\{ \left\lceil \frac{n}{r + 1} \right\rceil, \left\lceil \frac{k}{r} \right\rceil \right\}.$$

The proof of this theorem follows that of Theorem 1. It is known that, for fixed q , if n and L go to infinity, then $\frac{k_L^{(q)}(n, \sigma n)}{n} \rightarrow 1 - H_q(\sigma)$. Therefore, the rate R of an locally recoverable $(\sigma n, L)$ -list decodable code must satisfy,

$$\begin{aligned} R &\leq \min_{0 \leq x \leq \frac{r}{r+1}} x + (1 - x(1 + 1/r)) \left(1 - H_q \left(\frac{\delta}{1 - x(1 + 1/r)} \right) \right) \\ &= 1 - \max_{0 \leq x \leq \frac{r}{r+1}} \left[\frac{x}{r} + (1 - x(1 + 1/r)) H_q \left(\frac{\delta}{1 - x(1 + 1/r)} \right) \right]. \end{aligned}$$

REFERENCES

- [1] A. Barg, J. Justesen, and C. Thomsen. Concatenated codes with fixed inner code and random outer code. *IEEE Transactions on Information Theory*, 47(1):361, 2001.
- [2] A. Barg, A. Mazumdar, and G. Zémor. Weight distribution and decoding of codes on hypergraphs. *Advances in Mathematics of Communications (AMC)*, 2(4):433–450, 2008.
- [3] J. Boutros, O. Pothier, and G. Zemor. Generalized low density (tanner) codes. In *Communications, 1999. ICC'99. 1999 IEEE International Conference on*, volume 1, pages 441–445. IEEE, 1999.
- [4] V. R. Cadambe and A. Mazumdar. An upper bound on the size of locally recoverable codes. *CoRR*, abs/1308.3200, 2013.
- [5] A. Datta and F. E. Oggier. An overview of codes tailor-made for networked distributed data storage. *CoRR*, abs/1109.2317, 2011. <http://arxiv.org/abs/1109.2317>.
- [6] G. D. Forney. *Concatenated codes*. MIT, 1966.
- [7] R. Gallager. Low-density parity-check codes. *IRE Trans. Inform. Theory*, 8(1):21–28, Jan. 1962.
- [8] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin. On the locality of codeword symbols. *IEEE Transactions on Information Theory*, 58(11):6925–6934, nov. 2012.
- [9] S. Goparaju and R. Calderbank. Binary cyclic codes that are locally repairable. In *2014 IEEE International Symposium on Information Theory (ISIT)*, pages 676–680. IEEE, 2014.

- [10] C. Huang, M. Chen, and J. Li. Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems. In *Sixth IEEE International Symposium on Network Computing and Applications (NCA), 2007*, pages 79–86. IEEE, 2007.
- [11] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin. Erasure coding in windows azure storage. In *USENIX Annual Technical Conference (USENIX ATC), 2012*.
- [12] G. M. Kamath, N. Prakash, V. Lalitha, and P. V. Kumar. Codes with local regeneration. *CoRR*, abs/1211.1932, 2012. <http://arxiv.org/abs/1211.1932>.
- [13] G. M. Kamath, N. Prakash, V. Lalitha, P. V. Kumar, N. Silberstein, A. S. Rawat, O. O. Koyluoglu, and S. Vishwanath. Explicit mbr all-symbol locality codes. *arXiv preprint arXiv:1302.0744*, 2013.
- [14] M. Lentmaier and K. S. Zigangirov. On generalized low-density parity-check codes based on hamming component codes. *Communications Letters, IEEE*, 3(8):248–250, 1999.
- [15] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1977.
- [16] A. Mazumdar, V. Chandar, and G. W. Wornell. Local recovery properties of capacity-achieving codes. *Information Theory and Applications (ITA), San Diego*, Feb 2013.
- [17] F. Oggier and A. Datta. Self-repairing homomorphic codes for distributed storage systems. In *2011 Proceedings IEEE INFOCOM*, pages 1215–1223. IEEE, 2011.
- [18] D. S. Papailiopoulos and A. G. Dimakis. Locally repairable codes. In *Proceedings of 2012 IEEE International Symposium on Information Theory (ISIT)*, pages 2771–2775. IEEE, 2012.
- [19] D. S. Papailiopoulos, J. Luo, A. G. Dimakis, C. Huang, and J. Li. Simple regenerating codes: Network coding for cloud storage. In *2012 Proceedings IEEE INFOCOM*, pages 2801–2805. IEEE, 2012.
- [20] N. Prakash, G. M. Kamath, V. Lalitha, and P. V. Kumar. Optimal linear codes with a local-error-correction property. In *2012 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 2776–2780, 2012.
- [21] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath. Optimal locally repairable and secure codes for distributed storage systems. *arXiv preprint arXiv:1210.6954*, 2012.
- [22] A. S. Rawat, A. Mazumdar, and S. Vishwanath. On cooperative local repair in distributed storage. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–5. IEEE, 2014.
- [23] R. Roth. *Introduction to coding theory*. Cambridge University Press, 2006.
- [24] M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. G. Dimakis, R. Vadali, S. Chen, and D. Borthakur. Xoring elephants: Novel erasure codes for big data. *arXiv preprint arXiv:1301.3791*, 2013.
- [25] N. Silberstein and A. Zeh. Optimal binary locally repairable codes via anticode. *CoRR*, abs/1501.07114, 2015. <http://arxiv.org/abs/1501.07114>.
- [26] I. Tamo and A. Barg. A family of optimal locally recoverable codes. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 686–690. IEEE, 2014.
- [27] I. Tamo, D. S. Papailiopoulos, and A. G. Dimakis. Optimal locally repairable codes and connections to matroid theory. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1814–1818. IEEE, 2013.
- [28] A. Zeh and E. Yaakobi. Optimal linear and cyclic locally repairable codes over small fields. *arXiv preprint arXiv:1502.06809*, 2015.

APPENDIX

A. Proof of Lemma 1

Consider an r -locally recoverable (n, k, d) -code. For any $i \in \{1, 2, \dots, n\}$, let \mathcal{R}_i denote the corresponding repair-set; by definition $|\mathcal{R}_i| = r$. The key idea is to construct a set \mathcal{I} having the desired properties. Note that the proof is trivial for $t = 1$ since any codeword symbol in combination with the r symbols that form its local repair set form a valid choice of the set \mathcal{I} . The construction of the set \mathcal{I} is more challenging for $t > 1$. Our construction is essentially similar to [18]; we describe our construction here for completeness. We choose

$$\mathcal{I} = \left(\bigcup_{l=1}^t \{a_l\} \cup \mathcal{R}_{a_l} \cup \mathcal{S}_l \right)$$

where $a_1, a_2, \dots, a_t \in \{1, 2, \dots, n\}$ and $\mathcal{S}_l \subset \{1, 2, \dots, n\}, l = 1, 2, \dots, t$ are chosen as follows:

Begin Choose a_1 arbitrarily from $\{1, 2, \dots, n\}$. Choose \mathcal{S}_1 to be the null set.

Loop For $m = 2$ to $m = t$

Step 1: Choose a_m so that

$$a_m \notin \bigcup_{l=1}^{m-1} \{a_l\} \cup \mathcal{R}_{a_l} \cup \mathcal{S}_l$$

Step 2: Let $\mathcal{I}_{m-1} = \bigcup_{l=1}^{m-1} \{a_l\} \cup \mathcal{R}_{a_l} \cup \mathcal{S}_l$. Choose \mathcal{S}_m to be set of $m(r+1) - |\{a_m\} \cup \mathcal{R}_{a_m} \cup \mathcal{I}_{m-1}|$ elements, arbitrarily from $\{1, 2, \dots, n\} - \{a_m\} \cup \mathcal{R}_{a_m} \cup \mathcal{I}_{m-1}$.

End

This completes the construction. Note that \mathcal{I} constructed above has cardinality $t(r+1)$. It remains to show that $H(\mathcal{I}) \leq tr$. We now intend to show that $H(\mathcal{I}) = H(\mathcal{I} - \{a_1, a_2, \dots, a_t\})$ from which the desired bound would follow because of

$$H(\mathcal{I}) = H(\mathcal{I} - \{a_1, a_2, \dots, a_t\}) \leq t(r+1) - t = tr,$$

where we have used the fact that $H(\mathcal{A}) \leq |\mathcal{A}|$ for any set \mathcal{A} . We therefore intend to show a one-to-one mapping between $\{\mathbf{X}_{\mathcal{I} - \{a_1, a_2, \dots, a_t\}}\}$ and $\{\mathbf{X}_{\mathcal{I}}\}$. In other words, suppose that $\mathbf{X}_{\mathcal{I}} \neq \hat{\mathbf{X}}_{\mathcal{I}}$,

we need to prove that $\mathbf{X}_{\mathcal{I}-\{a_1, a_2, \dots, a_t\}} \neq \hat{\mathbf{X}}_{\mathcal{I}-\{a_1, a_2, \dots, a_t\}}$. Equivalently, suppose that $\mathbf{X}_{\{a_1, a_2, \dots, a_t\}} \neq \hat{\mathbf{X}}_{\{a_1, a_2, \dots, a_t\}}$, we need to prove that $\mathbf{X}_{\mathcal{I}-\{a_1, a_2, \dots, a_t\}} \neq \hat{\mathbf{X}}_{\mathcal{I}-\{a_1, a_2, \dots, a_t\}}$. Suppose a contradiction, i.e., suppose that $\exists, \mathbf{X}, \hat{\mathbf{X}} \in \mathcal{C}$ such that

$$\begin{aligned} \mathbf{X}_{\{a_1, a_2, \dots, a_t\}} &\neq \hat{\mathbf{X}}_{\{a_1, a_2, \dots, a_t\}} \\ \mathbf{X}_{\mathcal{I}-\{a_1, a_2, \dots, a_t\}} &= \hat{\mathbf{X}}_{\mathcal{I}-\{a_1, a_2, \dots, a_t\}} \end{aligned}$$

Define $\mathcal{B} = \{j : \mathbf{X}_j \neq \hat{\mathbf{X}}_j, j \in \{a_1, a_2, \dots, a_t\}\}$. Note that $\mathcal{B} \subseteq \{a_1, a_2, \dots, a_t\}$. Because of the definition of locality and because $\mathcal{R}_{a_i} \in \mathcal{I}$, the above conditions imply that

$$\mathcal{R}_i \cap \mathcal{B} \neq \phi, \forall i \in \mathcal{B} \quad (10)$$

In other words, the repair set associated with any element, i , in \mathcal{B} should have at least one element in \mathcal{B} , because $\mathbf{X}_j = \hat{\mathbf{X}}_j$ for all $i \neq j, j \in \mathcal{I} - \mathcal{B}$. We will show that this is a contradiction to our construction. In particular, we will throw away elements from \mathcal{B} one at a time to obtain, from (10), a relation of the form $j \cap \mathcal{R}_j \neq \phi$ for some $j \in \mathcal{B}$, which is a contradiction. To keep the notation clean, we will show the proof for $\mathcal{B} = \{a_1, a_2, \dots, a_m\}$, where $m = |\mathcal{B}|$. Our idea generalizes for arbitrary \mathcal{B} . By construction (Step 1), note that $a_m \notin \mathcal{R}_{a_i}, i = 1, 2, \dots, m-1$. Therefore, a_m is not a member of the repair sets of any of the elements of \mathcal{B} , and (10) implies that

$$\mathcal{R}_i \cap \{a_1, a_2, \dots, a_{m-1}\} \neq \phi, \forall i \in \{a_1, a_2, \dots, a_{m-1}\}$$

Similarly, note that $a_{m-1} \notin \mathcal{R}_{a_i}, i = 1, 2, \dots, m-2$ and Therefore, a_{m-1} is not a member of the repair sets of any of the elements of $\mathcal{B} - \{a_m\}$. So we get,

$$\mathcal{R}_i \cap \{a_1, a_2, \dots, a_{m-2}\} \neq \phi, \forall i \in \{a_1, a_2, \dots, a_{m-2}\}$$

Repeating the above procedure $m-1$ times, we get

$$\mathcal{R}_{a_1} \cap \{a_1\} \neq \phi,$$

which is a contradiction.

B. Proof of Lemma 2

Without loss of generality, let us assume that $\mathcal{I} = \{1, 2, \dots, |\mathcal{I}|\}$. Consider any element \mathbf{Z} of the $\mathcal{S} = \{\mathbf{X}_{\mathcal{I}} : \mathbf{X} \in \mathcal{C}\}$. Now, notice that the set of all elements of \mathcal{C} which have \mathbf{Z} as a “prefix” can be used to construct a codebook $\tilde{\mathcal{C}}(\mathbf{Z})$ of length $(n - |\mathcal{I}|)$. In particular denote

$$\tilde{\mathcal{C}}(\mathbf{Z}) = \{\mathbf{X}_{\{|\mathcal{I}|+1, |\mathcal{I}|+2, \dots, n\}} : \mathbf{X}_{\mathcal{I}} = \mathbf{Z}\}$$

In addition, we can deduce that the codebook $\tilde{\mathcal{C}}(\mathbf{Z})$, has minimum distance d . To see this, consider $\mathbf{U}, \mathbf{V} \in \tilde{\mathcal{C}}(\mathbf{Z})$ and note that

$$\Delta_H(\mathbf{U}, \mathbf{V}) = \Delta_H((\mathbf{Z}, \mathbf{U}), (\mathbf{Z}, \mathbf{V})) \geq d \quad (11)$$

where, above we have used the fact that, by definition of $\tilde{\mathcal{C}}(\mathbf{Z})$, the tuples (\mathbf{Z}, \mathbf{U}) and (\mathbf{Z}, \mathbf{V}) are elements of \mathcal{C} and therefore have a Hamming distance larger than or equal to d . Now, all we need to show is that there exists at least one $\hat{\mathbf{Z}} \in \mathcal{S}$ such that the dimension of $\tilde{\mathcal{C}}(\hat{\mathbf{Z}})$ is (at least) as large as $k - m$. This can be shown using an elementary probabilistic counting argument. Specifically, by assuming that \mathbf{Z} is uniformly distributed over \mathcal{S} , the average value of $|\tilde{\mathcal{C}}(\mathbf{Z})|$ can be bounded as follows.

$$\begin{aligned} |\mathcal{C}| = |\mathcal{Q}|^k &= \sum_{\mathbf{Z} \in \mathcal{S}} |\tilde{\mathcal{C}}(\mathbf{Z})| \\ &= |\mathcal{S}| E \left[|\tilde{\mathcal{C}}(\mathbf{Z})| \right] \\ \Rightarrow E \left[|\tilde{\mathcal{C}}(\mathbf{Z})| \right] &= \frac{|\mathcal{Q}|^k}{|\mathcal{S}|} \\ &\geq \frac{|\mathcal{Q}|^k}{|\mathcal{Q}|^m} = |\mathcal{Q}|^{k-m} \end{aligned}$$

where, above we have used the premise of the lemma, namely $|\mathcal{S}| = |\mathcal{Q}|^{H(\mathcal{I})} \leq |\mathcal{Q}|^m$. Therefore, there is at least one $\hat{\mathbf{Z}} \in \mathcal{S}$ such that $|\tilde{\mathcal{C}}(\hat{\mathbf{Z}})| \geq |\mathcal{Q}|^{k-m}$ thereby resulting in a $(n - |\mathcal{I}|, k - m, d)$ codebook over \mathcal{Q} . This completes the proof.

C. Proof of Theorem 2

For any two randomly chosen codewords, let W be the distance between them. The distribution of W is exactly the same as provided in the proof of Theorem 3.

More precisely, define for two codewords \mathbf{X}_i and \mathbf{X}_j , $1 \leq i, j \leq M, i \neq j$, the event $\Omega_{i,j} = \{\Delta_H(\mathbf{X}_i, \mathbf{X}_j) < d\}$. We have,

$$\Pr(\Omega_{i,j}) = \Pr(W < d).$$

Consider the dependency graph of the events $\{\Omega_{i,j}\}_{1 \leq i, j \leq M, i \neq j}$. In this graph two vertices corresponding to the events will have an edge between them if the events are dependent. This graph has order $M(M-1)$ and degree at most $2(M-1)$. Hence, using Lovász Local Lemma,

$$\Pr(\cap_{i,j} \bar{\Omega}_{i,j}) > 0,$$

as long as $\Pr(W < d)(2M-1) < \frac{1}{e}$. Note that, this means the existence of a locally recoverable code (with parameters n, k, d and r) in the ensemble as long as $\Pr(W < d) < \frac{1}{e(2M-1)}$. Plugging in the value calculated for $\Pr(W < d)$ from the proof of Theorem 3, we arrive at the statement of the theorem.