View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CWI's Institutional Repository

Construction of Optimal Locally Recoverable Codes and Connection with Hypergraph

Chaoping Xing
School of Electronics, Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai, 200240, P. R. China

Chen Yuan
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
chenyuan@cwi.nl

Abstract

Locally recoverable codes are a class of block codes with an additional property called locality. A locally recoverable code with locality r can recover a symbol by reading at most r other symbols. Recently, it was discovered by several authors that a q-ary optimal locally recoverable code, i.e., a locally recoverable code achieving the Singleton-type bound, can have length much bigger than $q+1$. In this paper, we present both the upper bound and the lower bound on the length of optimal locally recoverable codes. Our lower bound improves the best known result in [12] for all distance $d \geq 7$. This result is built on the observation of the parity-check matrix equipped with the Vandermonde structure. It turns out that a parity-check matrix with the Vandermonde structure produces an optimal locally recoverable code if it satisfies a certain expansion property for subsets of $\mathbb{F}_q$. To our surprise, this expansion property is then shown to be equivalent to a well-studied problem in extremal graph theory. Our upper bound is derived by an refined analysis of the arguments of Theorem 3.3 in [6].

2012 ACM Subject Classification Theory of computation → Error-correcting codes

Keywords and phrases Locally Repairable Codes, Hypergraph

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.98

Category Track A: Algorithms, Complexity and Games

Funding *Chen Yuan*: This research is supported by the European Union Horizon 2020 research and innovation programme under grant agreement No. 74079 (ALGSTRONGCRYPTO).

Acknowledgements We sincerely thank Prof. J. Verstraëte for his linking our condition (8) with the problem in extremal graph theory. He also provided us some references for latest results on extremal graph theory. We would also like to express our great gratitude to Profs. V. Guruswami, Q. Xiang and M. Lu for discussions and help.

1 Introduction

Motivated by applications in distributed and cloud storage systems, locally recoverable codes have been studied extensively in recent years. Informally speaking, a locally recoverable code (LRC for short) is a block code with an additional property called locality. For a locally recoverable code C of length n, dimension k and locality r, it was shown in [4] that the minimum distance $d(C)$ of C is upper bounded by

$$d(C) \leqslant n-k-\left\lceil \frac{k}{r} \right\rceil+2. \tag{1}$$

The bound (1) is called the Singleton-type bound for locally recoverable codes. A code achieving the above bound is usually called optimal.

© Chaoping Xing and Chen Yuan;
licensed under Creative Commons License CC-BY
46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 98; pp. 98:1–98:13
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1.1 Known results

Construction of optimal locally recoverable codes, i.e., block codes achieving the bound (1) is of both theoretical interest and practical importance. This is a challenging task and has attracted great attention in the last few years. In the literature, there are a few constructions available and some classes of optimal locally recoverable codes are known. A class of codes constructed earlier and known as pyramid codes [8] are shown to be codes that are optimal. In [14], Silberstein *et al* proposed a two-level construction based on the Gabidulin codes combined with a single parity-check $(r+1, r)$ code. Another construction [16] used two layers of MDS codes, a Reed-Solomon code and a special $(r+1, r)$ MDS code. A common shortcoming of these constructions relates to the size of the code alphabet which in all the papers is an exponential function of the code length, complicating the implementation. There was an earlier construction of optimal locally recoverable codes given in [13] with alphabet size comparable to code length. However, the construction in [13] only produces a specific value of the length n, i.e., $n = \lceil \frac{k}{r} \rceil (r+1)$. Thus, the rate of the code is very close to 1. There are also some existence results given in [13] and [15] with less restrictions on the locality r. But both results require large alphabet which is an exponential function of the code length.

A recent breakthrough construction was given in [15]. This construction naturally generalizes Reed-Solomon construction which relies on the alphabet of cardinality comparable to the code length n. The idea behind the construction is very nice. The only shortcoming of this construction is restriction on the locality r. Namely, $r+1$ must be a divisor of either $q-1$ or q, or $r+1$ is equal to a product of a divisor of $q-1$ and a divisor of q for certain q, where q is the code alphabet. This construction was extended via automorphism group of rational function fields by Jin, Ma and Xing [10] and it turns out that there is more flexibility on the locality and the code length can be $q+1$. For some particular locality such as $r = 2, 3, 5, 7, 11$ or 23, it was shown that there exist q-ary optimal locally recoverable codes with length up to $q + 2\sqrt{q}$ via elliptic curves [11]. All these results are aimed at the optimal LRC with large distance.

Unlike classical MDS codes, it is surprising to discover that the optimal LRCs can have super-linear code length in alphabet size q. Barg et.al, [1] gave optimal LRCs by using algebraic surfaces of length $n \approx q^2$ when the distance $d = 3$ and $r \leqslant 4$. This inspired the construction of the optimal LRC with unbounded length and distance $d = 3, 4$ [12]. Furthermore, it was shown in [6] that an optimal LRC with $d \geq 5$ must have length upper bounded in terms of alphabet size q. More precisely, they showed that the length of an optimal q-ary linear LRC with distance $d \geqslant 5$ and locality r is upper bonded by $O\left(dq^{3+\frac{4}{d-4}}\right)$. As for the lower bound, they presented an explicit construction of optimal LRCs with code length $\Omega_r\left(q^{1+\frac{1}{\lfloor (d-3)/2 \rfloor}}\right)$ provided that $d \leq r+2$, where Ω_r means that the implied constant depends on r. One can see that there is still a huge gap between the lower bound and the upper bound. Following this discovery, there are several works dedicated to constructing the maximum length of optimal LRCs. The paper [9] aimed at the optimal LRC with small distance $d = 5$ or 6. In particular, for $d = 6$, the results given in [9] are obtained subject to the constraint that q is even.

1.2 Our results, comparisons and a conjecture

The main result of this paper can be summarized as follows.

► **Theorem 1.** *Suppose that $r \geqslant d-2$ and $(r+1)|n$. Then*

(i) *there exists an explicit construction of optimal locally recoverable codes with length $n = q^{2-o(1)}$, minimum distance d and locality r for $d = 7, 8$;*

(ii) *there exists an explicit construction of optimal locally recoverable codes with length* $n = q^{\frac{3}{2}-o(1)}$*, minimum distance* d *and locality* r *for* $d = 9, 10$*;*

(iii) *there exist optimal locally recoverable codes with length* $n = \Omega_{r,d}\left(q(q\log q)^{\frac{1}{\lfloor (d-3)/2\rfloor}}\right)$*, minimum distance* d *and locality* r *for* $d \geq 11$*; and*

(iv) *there exists an explicit construction of optimal locally recoverable code with length* $n = \Omega_{r,d}\left(q^{1+\frac{1}{\lfloor (d-3)/2\rfloor}}\right)$*, minimum distance* d *and locality* r *for a constant* $d \geq 11$*. Moreover, the complexity of this construction is upper bounded by* $O(n^d)$*.*

The first three results are derived from extremal graph theory (see Section 5). The last one is derived from the probabilistic arguments (see Section 4).

The first two results improve on the result in [6] which only achieves $n = \Omega(q^{3/2})$ for $d = 7, 8$ and $n = \Omega(q^{4/3})$ for $d = 9, 10$. The third one outperforms the result in [6] by a $(\log q)^{\frac{1}{\lfloor (d-3)/2\rfloor}}$ multiplicative factor. In addition, for $d = 6$, we are able to remove the constraint required in [9] that q is even.

Although it was proved in [6] that the length of an optimal locally recoverable code is upper bounded by $q^{3+O\left(\frac{1}{d}\right)}$, both the constructions in [6] and this paper show from different angles that the length of an optimal locally recoverable code only achieve $q^{1+O\left(\frac{1}{d}\right)}$. Furthermore, via an upper bound from extremal graph theory, our construction in this paper can achieve at most $O\left(q^{1+\frac{2}{\lfloor (d-1)/2\rfloor}}\right)$ (see Section 5). Thus, we make the following conjecture.

▶ **Conjecture 2.** *Every optimal locally recoverable code with minimum distance* d *and locality* r *has length upper bounded by* $q^{1+O\left(\frac{1}{d}\right)}$*.*

In addition to the above lower bound on length of optimal locally recoverable codes, we also provide an improved upper bound by refining the analysis of the arguments of Theorem 3.3 in [6].

▶ **Theorem 3** (Informal)**.** *Let* C *be an optimal* $[n, k, d]_q$*-linear locally repairable code with the locality* r*. If* $d \geqslant 5$*, then*

$$n \leq \begin{cases} O(q^3) & \text{if } d \bmod r+1 > 5 \text{ or } d \bmod r+1 < 2, \\ O(q^2) & \text{if } 2 \leq d \bmod r+1 \leq 5. \end{cases} \tag{2}$$

1.3 Our techniques

For minimum distance $d \geq 7$, the only optimal locally recoverable code with super-linear code length was given in [6]. In this paper, we present another construction for optimal LRCs for $d \geqslant 5$. Our idea comes from generalized Reed-Solomon codes where parity-check matrices have the Vandermonde structure. This idea was already employed in [9] for $d = 5, 6$. Similar to [9], we divide a parity-check matrix into disjoint blocks, each block with $r + 1$ columns. We require that each block of this matrix has a Vandermonde matrix structure. In order that the parity-check matrix with this structure produces an optimal locally recoverable code, elements in these blocks must satisfy certain expansion property. This property allows us to relate optimality of a locally recoverable code to a well-studied problem in extremal graph theory. With the help of extremal graph theory, we succeed to improve all of the best known results in [6] for $d \geq 7$.

Furthermore, by a random or probabilistic argument, we show an existence result. Moreover, for constant d the probabilistic method for the existence result can be converted into a deterministic algorithm via method of conditional probabilities. Thus, we obtain an algorithmic construction in polynomial time, i.e., Theorem 1(iv). The result of Theorem 1(iv) matches the result given in [6]. However, our parity-check matrix is more structured and this may lead to some other applications.

1.4 Organization

The paper is organized as follows. In Section 2, we briefly introduce locally recoverable codes and some basic notations on graph theory. Section 3 presents a necessary and sufficient condition for which a Vandermonde-type parity-check matrix produces an optimal locally recoverable code in terms of certain expansion property for subsets of $\mathbb{F}_q$. In Section 4, we first show an existence result via a probabilistic method. Then this probabilistic method is converted into an algorithmic construction in polynomial time. In Section 5, we show that the necessary and sufficient condition derived in Section 2 is equivalent to a central problem in extremal graph theory. By applying the known results from extremal graph theory, we obtain the desired results. Finally, in Section 6, we prove a general upper bound on the optimal LRC.

2 Preliminaries

2.1 Locally recoverable codes

Let q be a prime power and $\mathbb{F}_q$ be the finite field with q elements and denote by $[n]$ the set $\{1, 2, \dots, n\}$. In this paper, we consider linear locally recoverable codes only. An $[n, k, d]$ linear code C is a k-dimensional subspace of $\mathbb{F}_q^n$ with minimum (Hamming) distance d. The (Euclidean) dual code of C, denoted by $C^\perp$, is defined by $C^\perp = \{\mathbf{b} \in \mathbb{F}_q^n : \mathbf{c} \cdot \mathbf{b} = 0 \text{ for all } \mathbf{c} \in C\}$, where $\mathbf{c} \cdot \mathbf{b}$ denotes the standard inner product of the two vectors $\mathbf{b}$ and $\mathbf{c}$.

Informally speaking, a block code is said to have locality r if every coordinate of a given codeword can be recovered by accessing at most r other coordinates of this codeword. There are several equivalent definitions of locally recoverable codes. A formal definition of a locally recoverable code with locality r is given as follows.

▶ **Definition 4.** A q-ary block code C of length n is called a locally recoverable code or locally repairable code (LRC for short) with locality r if for any $i \in [n]$, there exists a subset $R_i \subseteq [n] \setminus \{i\}$ of size r such that for any $\mathbf{c} = (c_1, \dots, c_n) \in C$, c_i can be recovered by $\{c_j\}_{j \in R_i}$, i.e., for any $i \in [n]$, there exists a subset $R_i \subseteq [n] \setminus \{i\}$ of size r such that for any $\mathbf{u}, \mathbf{v} \in C$, $\mathbf{u}_{R_i \cup \{i\}} = \mathbf{v}_{R_i \cup \{i\}}$ if and only if $\mathbf{u}_{R_i} = \mathbf{v}_{R_i}$. The set R_i is called a recovering set of i.

In literature, there are various definitions for locally recoverable code and all of them are equivalent. For example, we have the following two definitions that are equivalent to Definition 4. For the sake of completeness, we give a proof.

▶ **Lemma 5.** *A q-ary code C of length n is a locally recoverable code if and only if one of the followings holds.*

(i) *For any $i \in [n]$, there exists a subset $R_i \subseteq [n] \setminus \{i\}$ of size r such that position i of every codeword $\mathbf{c} \in C$ is determined by $\mathbf{c}_{R_i}$, i.e, there is a function $f_i(x_1, \dots, x_r)$ (independent of $\mathbf{c}$ and only dependent on i) such that $c_i = f_i(\mathbf{c}_{R_i})$, where $\mathbf{c}_{R_i}$ stands for the projection of $\mathbf{c}$ at R_i.*

(ii) *For any $i \in [n]$, there exists a subset $R_i \subseteq [n] \setminus \{i\}$ of size r such that*

$$C_{R_i}(i, \alpha) \cap C_{R_i}(i, \beta) = \emptyset$$

for any $\alpha \neq \beta \in \mathbb{F}_q$, where $C(i, \alpha) = \{\mathbf{c} \in C : c_i = \alpha\}$ and $C_{R_i}(i, \alpha)$ denotes the projection of $C(i, \alpha)$ on R_i.

The Singleton (upper) bound in (1) is given in terms of minimum distance d. We can also rewrite this bound in terms of dimension k.

▶ **Lemma 6.** *Let n, k, d, r be positive integers with $(r+1)|n$. If the Singleton-type bound* (1) *is achieved, then*

$$n-k=\frac{n}{r+1}+d-2-\left\lfloor\frac{d-2}{r+1}\right\rfloor. \tag{3}$$

Conversely, if $d-2 \not\equiv r \pmod{r+1}$ and the equlity (3) *is satisfied, then the Singleton-type bound* (1) *is achieved.*

The proof is straightforward and can be found in [6].

▶ Remark 7. If $d-2 \equiv r \pmod{r+1}$, one can verify that (3) implies that $r|k$. In this case, by [4, Corollary 10] one cannot achieve the Singleton-type bound (1) with equality and one must have $d \leqslant n-k-\lceil\frac{k}{r}\rceil+1$. Therefore in this case we say an LRC attaining this latter bound as optimal.

▶ **Corollary 8.** *If $r \geqslant d-2$, then an $[n, k, d]$ locally recoverable code with locality r is optimal if*

$$n-k-\frac{n}{r+1}=d-2. \tag{4}$$

Proof. As $r \geqslant d-2$, $\left\lfloor\frac{d-2}{r+1}\right\rfloor=0$. Hence, (3) and (4) are equivalent. ◀

The locality of a locally recoverable code C can be determined by a parity-check matrix of C as follows. Assume that $(r+1)|n$. Let $m=\frac{n}{r+1}$ and let D_i be $(n-k-m)\times(r+1)$ matrices. Put

$$H=\left(\begin{array}{c|c|c|c}
\mathbf{1} & \mathbf{0} & \cdots & \mathbf{0}\\ \hline
\mathbf{0} & \mathbf{1} & \cdots & \mathbf{0}\\ \hline
\vdots & \vdots & \ddots & \vdots\\ \hline
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}\\ \hline
D_1 & D_2 & \cdots & D_m
\end{array}\right), \tag{5}$$

where $\mathbf{1}$ and $\mathbf{0}$ stand for the all-one row vector and the zero row vector of length $r+1$, respectively. Let C be the code with H as a parity-check matrix. Then it is clear that the dimension of C is at least k. Furthermore, we claim that the locality of C is r. Indeed, let $\mathbf{c}=(c_1, c_2, \dots, c_n)$ be a codeword of C, then $\sum_{j=1+(r+1)i}^{(r+1)(i+1)} c_j=0$ for $0 \leqslant i \leqslant m-1$ as $H\mathbf{c}^T=\mathbf{0}$. Hence, a coordinate c_j with $j \in \{1+(r+1)i, \dots, (r+1)(i+1)\}$ for some $0 \leqslant i \leqslant m-1$ can be repaired by $\mathbf{c}_{R_j}$ with $R_j=\{1+(r+1)i, \dots, (r+1)(i+1)\} \setminus \{j\}$.

In conclusion, to see if a linear code C with a parity-check matrix H of the form (5) is an optimal locally recoverable code, it is sufficient to check if the minimum distance of C satisfies (4) for $r \geqslant d-2$.

2.2 Graphs

A undirected graph G is a pair $G=(V, E)$, where V is a finite set and E is a set consisting of some subsets of size 2 of V. An element of V is called a vertex and an element of E is called an edge. A subgraph G' of a graph G is a graph whose vertex set and edge set are subsets of those of G. We say that G has a cycle $(v_1, \dots, v_m)$ if $\{v_i, v_{i+1}\} \in E$ for $i=1, \dots, m-1$ and $\{v_m, v_1\} \in E$. The following Lemma 9 provides a simple but useful way to determine if G contains a cycle. The proof can be found in any textbook about graph theory (see [3] for instance).

▶ **Lemma 9.** *An undirected graph G contains a cycle if $|E| \geq |V|$.*

Apart from the above usual definition of graphs, we also require some results on hypergraph in this paper. A hypergraph is a generalization of a graph in which an edge can join any number of vertices. Formally, a hypergraph H is a pair $H = (X, E)$ where X is a set of elements called vertices, and E is a set of non-empty subsets of X called hyperedges or edges. Therefore, E is a subset of $2^X \setminus \{\emptyset\}$, where 2^X stands for the power set of X.

▶ **Definition 10** (r-uniform Hypergraph (or r-graph for short))**.** *A hypergraph $H = (X, E)$ is called r-uniform if every hyperedge in E has size r. In other words, every hyperedge of an r-uniform hypergraph connects exactly r vertices.*

There are several ways to define cycles in a hypergraph that coincide with the definition of cycles in the usual graph. In this paper, we use the Berge cycle as the generalization of cycles in the usual graph.

▶ **Definition 11** (Berge cycle)**.** A r-uniform hypergraph $H = (X, E)$ contains a Berge k-cycle $(v_1, \ldots, v_k)$ if there exist k hyperedges $e_1, \ldots, e_k \in E$ such that $\{v_{i-1}, v_i\} \subseteq e_i$ for $i = 2, \ldots, k$ and $\{v_1, v_k\} \subseteq e_1$.

3 A criterion on minimum distance

It follows from Corollary 8 that for $d \leq r+2$, a locally recoverable code with parity-check matrix H in (5) is optimal provided that any $d-1$ columns of H are linearly independent and each D_i is a $(d-2) \times (r+1)$ matrix.

Let $\mathbb{F}_q$ be a finite field and put $m = \frac{n}{r+1}$. Assume that $A_1, \ldots, A_m$ are subsets of $\mathbb{F}_q$, each of size $r+1$. Let $A_i = \{a_{i,1}, \ldots, a_{i,r+1}\}$ for $i = 1, \ldots, m$. Let $\mathbf{a}_{i,j} = (a_{i,j}, a_{i,j}^2, \ldots, a_{i,j}^{d-2})$ and put $D_i = (\mathbf{a}_{i,1}^T, \mathbf{a}_{i,2}^T, \ldots, \mathbf{a}_{i,r+1}^T)$. Thus, D_i is a Vandermonde-type matrix. Let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be the standard basis of vector space $\mathbb{F}_q^m$, i.e., all components of $\mathbf{e}_i$ are 0 except that the i-th component is 1. Then, we can rewrite H as follow.

$$H = \begin{pmatrix} \mathbf{e}_1^T & \cdots & \mathbf{e}_1^T & \cdots & \mathbf{e}_m^T & \cdots & \mathbf{e}_m^T \\ \mathbf{a}_{1,1}^T & \cdots & \mathbf{a}_{1,r+1}^T & \cdots & \mathbf{a}_{m,1}^T & \cdots & \mathbf{a}_{m,r+1}^T \end{pmatrix}. \tag{6}$$

We now present a sufficient and necessary condition under which any $d-1$ columns of the matrix H in (6) are linearly independent.

▶ **Theorem 12.** *For $d \geqslant 5$, then any $d-1$ columns of H defined in* (6) *are linearly independent if and only if $|\bigcup_{i \in S} A_i| \geq r|S| + 1$ for any $S \subseteq [m]$ of size no more than $t = \lfloor \frac{d-1}{2} \rfloor$.*

Proof. We first prove the "if" direction. Let $\mathbf{h}_{i,j}$ be the (i,j)th column of H, i.e., $\mathbf{h}_{i,j} = (\mathbf{e}_i, \mathbf{a}_{i,j})^T$ for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant r+1$. Choose any $d-1$ columns $\{\mathbf{h}_{i,j}\}_{1 \leqslant i \leqslant m; j \in S_i}$ of H, where S_i are subsets of $[r+1]$ satisfying $\sum_{i=1}^m |S_i| = d-1$. Let $A_i' = \{a_{i,j} \in A_i : j \in S_i\}$, i.e., A_i' is the subset of A_i where each element is associated with one of the $d-1$ columns. Let H' be the $(n-k-m) \times (d-1)$ matrix consisting of these $d-1$ columns. We are going to show that H' has rank $d-1$. We assume that S_i is either empty or of size at least 2. Otherwise, the unique column selected from D_i with $|S_i| = 1$ must be linearly independent from the rest $d-2$ columns. We can consider the linear independence of the rest $d-2$ columns instead. Now, we assume that there are at most t non-empty sets S_i. Let $A = \{a_{i,j}\}_{1 \leqslant i \leqslant m; j \in S_i}$.

Assume that $A=\{a_1,\dots,a_s\}$ has s distinct elements. If $s=d-1$, then by elementary row operations, one can find a $(d-1)\times(d-1)$ Vandermonde submatrix of the form

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{a}_1^T & \mathbf{a}_2^T & \cdots & \mathbf{a}_{d-1}^T \end{pmatrix}$$

of H', where $\mathbf{a}_i=(a_i,a_i^2,\dots,a_i^{d-2})$. Thus, the rank of H' is $d-1$.

We proceed to the case where $s<d-1$. By permuting the columns of H', we obtain a matrix of the following form:

$$H_1=\left(\begin{array}{cccc|ccc} \mathbf{e}_{i_1}^T & \mathbf{e}_{i_2}^T & \cdots & \mathbf{e}_{i_s}^T & \mathbf{e}_{i_{s+1}}^T & \cdots & \mathbf{e}_{i_{d-1}}^T \\ \mathbf{a}_1^T & \mathbf{a}_2^T & \cdots & \mathbf{a}_s^T & \mathbf{a}_{s+1}^T & \cdots & \mathbf{a}_{d-1}^T \end{array}\right),$$

where $1\leqslant i_1\leqslant i_2\leqslant\cdots\leqslant i_{d-1}\leqslant m$ and $\{a_{s+1},\dots,a_{d-1}\}$ is a subset of A. Thus, a_j belongs to A_{i_j} for $1\leqslant j\leqslant d-1$. By elementary column operations, we can erase $\mathbf{a}_{s+i}^T$ since it also appears in one of the first s columns. Hence, H_1 is equivalent to

$$H_2=\left(\begin{array}{cccc|ccc} \mathbf{e}_{i_1}^T & \mathbf{e}_{i_2}^T & \cdots & \mathbf{e}_{i_s}^T & \mathbf{e}_{i_{s+1}}^T-\mathbf{e}_{k_{s+1}}^T & \cdots & \mathbf{e}_{i_{d-1}}^T-\mathbf{e}_{k_{d-1}}^T \\ \mathbf{a}_1^T & \mathbf{a}_2^T & \cdots & \mathbf{a}_s^T & \mathbf{0}^T & \cdots & \mathbf{0}^T \end{array}\right),$$

where $\{k_{s+1},\dots,k_{d-1}\}$ is a subset of $\{i_1,\dots,i_s\}$. Since H_2 is an upper left triangular block matrix, showing that H_2 is a full-rank matrix is equivalent to showing both $(\mathbf{a}_1^T,\mathbf{a}_2^T,\dots,\mathbf{a}_s^T)$ and $(\mathbf{e}_{i_{s+1}}^T-\mathbf{e}_{k_{s+1}}^T,\cdots,\mathbf{e}_{i_{d-1}}^T-\mathbf{e}_{k_{d-1}}^T)$ have full rank. Note that $(\mathbf{a}_1^T,\mathbf{a}_2^T,\dots,\mathbf{a}_s^T)$ is a $(d-2)\times s$ Vandermonde matrix and hence it has full rank s. It remains to show that $\mathbf{e}_{i_{s+1}}-\mathbf{e}_{k_{s+1}},\dots,\mathbf{e}_{i_{d-1}}-\mathbf{e}_{k_{d-1}}$ are linearly independent. Suppose they were linearly dependent. Then there exist elements $\lambda_{s+1},\dots,\lambda_{d-1}\in\mathbb{F}_q$ which are not all zero such that

$$\sum_{j=s+1}^{d-1}\lambda_j(\mathbf{e}_{i_j}-\mathbf{e}_{k_j})=0.$$

Let P be the subset of $\{s+1,\dots,d-1\}$ such that $\lambda_i\neq 0$ if and only if $i\in P$. It follows that

$$\sum_{i\in P}\lambda_i(\mathbf{e}_{j_i}-\mathbf{e}_{k_i})=0. \tag{7}$$

Let $U=\{j_i: i\in P\}$, $V=\{k_i: i\in P\}$ and $W=U\cup V$. As both U and V are subsets of $\{i\in[m]:\ |S_i|\geqslant 2\}$, we have $|W|\leqslant t=\lfloor\frac{d-1}{2}\rfloor$. Since λ_i is nonzero for all $i\in P$, every $\ell\in W$ must appear at least twice in the multiset consisting of elements of U and V. Otherwise, $\mathbf{e}_\ell$ could not be cancelled in (7). This implies $|W|\le|P|$.

On the other hand, for each $a_i\in A$, there is exactly one subset A'_{k_i} containing a_i since the first s columns have s distinct $\mathbf{a}_i$. Furthermore, let $t_i=|\{\ell\in U: a_i\in A'_\ell\}|$. It follows that $\sum_{a_i\in A}t_i=|P|$ and a_i belongs to t_i+1 subsets in $\{A'_\ell:\ell\in W\}$. This implies

$$\left|\bigcup_{\ell\in W}A_\ell\right|\le\sum_{\ell\in W}|A_\ell|-\sum_{i=1}^{s}t_i=(r+1)|W|-|P|$$

since $A'_\ell\subseteq A_\ell$. Combining with the condition $|\bigcup_{\ell\in W}A_\ell|\ge r|W|+1$ forces $|W|\ge|P|+1$. A contradiction occurs and we complete the proof of the "if" direction.

We proceed to the "only if" direction. First, we claim that $|A_i\cap A_j|\le 1$ for any $i\neq j$. Otherwise, we may assume that $A_i\cap A_j$ contains two distinct elements a_1 and a_2. Thus, H contains the four linearly dependent columns $(\mathbf{e}_i,\mathbf{a}_1)^T,(\mathbf{e}_i,\mathbf{a}_2)^T,(\mathbf{e}_j,\mathbf{a}_1)^T$ and $(\mathbf{e}_j,\mathbf{a}_2)^T$.

We prove the "only if" part by contradiction. Without loss of generality, we assume that the first s subsets $A_1, \ldots, A_s$ do not satisfy the condition, i.e. $|\bigcup_{i=1}^{s} A_i| \leq sr$, where s satisfies $s \leqslant t$. Define an undirected graph $G = ([s], E)$ such that $\{i,j\} \in E$ if and only if $A_i \cap A_j \neq \emptyset$. By inclusion-exclusion principle, we have

$$rs \geqslant \left|\bigcup_{i=1}^{s} A_i\right| \geq \sum_{i=1}^{s} |A_i| - \sum_{(i,j)\in E} 1 = s(r+1) - |E|.$$

This implies $|E| \geq s$. By Lemma 9, there exists a cycle in this undirected graph. Without loss of generality, we may assume that $(1, \ldots, \ell)$ is a cycle, i.e., $\{i, i+1\} \in E$ for $i = 1, \ldots, \ell-1$ and $\{\ell, 1\} \in E$. By the definition of E, A_i and A_{i+1} contains a common element $\{a_{j_i}\}$. Then, we can pick two columns $(\mathbf{e}_i, \mathbf{a}_{j_{i-1}})^T$ [1] and $(\mathbf{e}_i, \mathbf{a}_{j_i})^T$ from the i-th block D_i for $i = 1, \ldots, \ell$. These 2ℓ columns are linearly dependent since

$$\sum_{i=1}^{\ell} \Big((\mathbf{e}_i, \mathbf{a}_{j_{i-1}}) - (\mathbf{e}_i, \mathbf{a}_{j_i}) \Big) = \sum_{i=1}^{\ell} (0, \mathbf{a}_{j_{i-1}} - \mathbf{a}_{j_i}) = \mathbf{0}.$$

The proof is completed. ◀

By Theorem 12, we immediately obtain the following result.

▶ **Theorem 13.** *If $t = \lfloor \frac{d-1}{2} \rfloor \geqslant 2$ and $(r+1)|n$, then there exists a q-ary optimal linear LRC with length n, minimum distance d and locality r provided that there are $m = \frac{n}{r+1}$ sets $A_1, \ldots, A_m \subseteq \mathbb{F}_q$ such that*

$$\begin{array}{ll} |A_i| = r+1 & \text{for } 1 \leq i \leq m, \\ |\bigcup_{i\in S} A_i| \geq |S|r+1 & \text{for any } S \subseteq [m] \text{ of size at most } t. \end{array} \tag{8}$$

▶ Remark 14. We point out that there is another way to look at (8) as one reviewer suggests. Define an unbalanced bipartite expander graph where the vertex u_i on the left hand side are associated with sets A_i and the vertex v_j on the right hand side are associated with an element a_j in $\mathbb{F}_q$. u_i and v_j are adjacent if and only if a_j is contained in A_i. The expansion property (8) is now equivalent to the existence of good unbalanced bipartite graph [5]. However, the parameters discussed in this paper are not in the scope of explicit construction of good unbalanced bipartite graph, i.e., the expansion property in (8) is too strong for all known explicit constructions.

4 Random and algorithmic constructions

In the previous section, we converted the construction of optimal LRCs into a problem of finding subsets of $\mathbb{F}_q$ satisfying (8). In this section, we first present a probabilistic construction of subsets satisfying (8). In addition, we can derandomize this probabilistic construction into a deterministic construction in polynomial time if d is constant. The case $t = 2$, i.e., $d = 5$ and 6, is equivalent to the design of constant weight codes [9]. In this section, we assume $t \geq 3$. Since the algebraic structure is not important for the union of set, we replace $\mathbb{F}_q$ with $[q]$ from now on.

▶ **Theorem 15.** *There exist $m = \left\lceil \frac{q^{1+\frac{1}{t-1}}}{2t^2(r+1)^{2+\frac{2}{t-1}}} \right\rceil$ sets $A_1, \ldots, A_m$ satisfying (8) provided q is large enough.*

[1] Define $\mathbf{a}_{j_0} = \mathbf{a}_{j_\ell}$ for simplicity.

Proof. Let $X_i = \{x_{i,1}, \dots, x_{i,r+1}\}, i = 1, \dots, 2m$ be the set picked uniformly at random over all $(r+1)$-sized subsets of $[q]$. Define the binary random variable Y_S such that $Y_S = 1$ if $|\bigcup_{i\in S} X_i| \le |S|r$ and 0 otherwise. Our goal is to bound the expectation $E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S\right]$. Without loss of generality, we may assume that $S = \{1, \dots, a\}$ for some $1 < a \le t$. We order the random variables in $X_i, i = 1, \dots, a$, i.e., $x_{1,1}, \dots, x_{1,r+1}, \dots, x_{a,1}, \dots, x_{a,r+1}$. We want to bound the probability of the event $Y_S = 1$, i.e., at least a elements repeated in this sequence. Given an element $x_{i,j}$, the probability that $x_{i,j} \ne x_{i',j'}$ for some $x_{i',j'}$ prior to $x_{i,j}$ is at least $1 - \frac{(i-1)(r+1)+j}{q} \ge 1 - \frac{a(r+1)}{q}$. Taking over all sets of size at least a in this sequence, the probability of $Y_S = 1$ is at most

$$\sum_{i=a}^{a(r+1)} \binom{a(r+1)}{i} \left(\frac{a(r+1)}{q}\right)^i \le \sum_{i=a}^{a(r+1)} \frac{(a(r+1))^i}{i!} \left(\frac{a(r+1)}{q}\right)^i \le \frac{1.1}{a!} \left(\frac{a^2(r+1)^2}{q}\right)^a.$$

for $q \ge 10a^2(r+1)^2$. It follows that

$$\begin{aligned} E\left[\sum_{S\subset[2m],|S|\le t} Y_S\right] &= \sum_{i=2}^{t} \sum_{S\subset[2m],|S|=i} \Pr[Y_S = 1] \le \sum_{i=2}^{t} \binom{2m}{i} \frac{1.1}{i!} \left(\frac{i^2(r+1)^2}{q}\right)^i \\ &\le \sum_{i=2}^{t} 1.1(\frac{1}{i!})^2 \left(\frac{2mi^2(r+1)^2}{q}\right)^i \le \sum_{i=2}^{t} 1.1 \left(\frac{1}{i!}\right)^2 \left(\frac{q}{(r+1)^2}\right)^{\frac{i}{t-1}} \\ &\le 1.1 \times 1.5 \left(\frac{1}{t!}\right)^2 \left(\frac{q}{(r+1)^2}\right)^{\frac{t}{t-1}} \le \frac{2}{4t^2} \left(\frac{q}{(r+1)^2}\right)^{\frac{t}{t-1}} \le m. \end{aligned}$$

for $q \ge t^{2t}3^t(r+1)$ and $t \ge 3$. The second inequality is due to $\binom{2m}{i} \le \frac{(2m)^i}{i!}$ and the third inequality is due to

$$\left(\frac{1}{i!}\right)^2 \left(\frac{q}{(r+1)^2}\right)^{\frac{i}{t-1}} \ge 3 \left(\frac{1}{(i-1)!}\right)^2 \left(\frac{q}{(r+1)^2}\right)^{\frac{i-1}{t-1}}.$$

That means there exists $2m$ $(r+1)$-sized sets $A_1, \dots, A_{2m}$ such that there are at most m subsets $S \subseteq [2m]$ with $|\bigcup_{i\in S} A_i| \le |S|r$. For each of these m subsets S, remove one set from $A_i, i \in S$. The desired result follows as we remove at most m sets. ◀

Theorem 15 is an existence proof. However, if t is a constant, it is possible to turn this argument into an algorithm via the method of conditional probabilities.

▶ **Theorem 16.** *There exists a polynomial-time deterministic algorithm to find m sets in Theorem 15 provided that t is a constant.*

Proof. We follow the same notation in Theorem 15. Let $X_i = \{x_{i,1}, \dots, x_{i,r+1}\}$ be a random set of size $r+1$. Our goal is to minimize $E[\sum_{S\subset[2m],|S|\le t} Y_S]$ by fixing the set X_i one by one. Since

$$\begin{aligned} E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S\right] &= \sum_{A\subset[q],|A|=r+1} E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S|X_1 = A\right] \Pr[X_1 = A] \\ &= \frac{1}{\binom{q}{r+1}} \sum_{A\subset[q],|A|=r+1} E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S|X_1 = A\right], \end{aligned}$$

there exists a set A such that $E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S|X_1 = A\right] \le E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S\right]$. If $r+1$ is a constant, we only need to enumerate all subsets of size $r+1$ in polynomial time. However, if $r+1$ is not a constant, we enumerate $x_{1,1} \in X_1$ instead of the whole set, i.e., minimizing

$E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S | x_{1,1}=a_{1,1}\right]$ for $a_{1,1}\in[q]$. Given a subset $S\subseteq[2m]$ of size t, let us show how to compute $E[Y_S|x_{1,1}=a_{1,1}]$. Without loss of generality, we assume $S=\{1,\dots,t\}$. We list $t(r+1)$ random elements $x_{1,1}=a_{1,1}, x_{1,2},\dots,x_{1,r+1},\dots,x_{t,1},\dots,x_{t,r+1}$. For large enough q, it suffices to approximate $E[Y_S|x_{1,1}=a_{1,1}]$ by counting the number of sequences where there are exact t repetitions. There are $\binom{(r+1)t}{t}$ combinations of these t positions. Let $R\subseteq[t]\times[r+1]$ be any set of size t representing the t positions. we first remove these t positions from the sequence. The remaining tr positions in the sequence must have distinct elements and there are $\prod_{i=0}^{rt-1}(q-i)$ ways to pick these tr elements. Assume that we assign $1,\dots,rt$ to these rt positions. To obtain our final result, we multiply it by $\prod_{i=0}^{rt-1}(q-i)$. It remains to fill our sequence by adding back the t positions in R. For each $(i,j)\in R$, we enumerate all possible choices of $x_{i,j},(i,j)\in R$ and find out the number of combinations that there are exact t repetitions in the resulting sequence. There are at most q^t ways to do the enumeration. Then, we obtain the exact value of $E[Y_S|x_{1,1}=a_{1,1}]$. Observe that there are at most $\sum_{i=2}^{t}\binom{n}{i}$ subsets S. Thus, this expectation can be computed in polynomial time as t is a constant. We do it $r+1$ times so as to fix all elements in X_1. Given $A_1,\dots,A_k$, our goal is to find $X_{k+1}=A_{k+1}$ to minimize the expectation

$$E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S|X_1=A_1,\dots,X_k=A_k\right]\le E\left[\sum_{S\subseteq[2m],|S|\le t} Y_S\right].$$

It can be done in the same way as $X_1,\dots,X_{k-1}$ are already fixed. After we fix all these $2m$ sets, we will obtain $A_1,\dots,A_{2m}$ with the same property as Theorem 15 claims. Then, we enumerate all t-sized subsets $S\subseteq[q]$ and do the same as Theorem 15 does. The resulting subsets are the output of our algorithm. The number of these subsets is at least m. Since t is constant, all this operation is done in polynomial time. The proof is completed. ◀

The following is a direct consequence of Theorem 12, Theorem 15 and Theorem 16.

▶ **Theorem 17.** *For $d\geqslant 5$, put $t=\left\lfloor\frac{d-1}{2}\right\rfloor$. If $r\geqslant d-2$, $(r+1)|n$ and q is sufficiently large, then there exists a q-ary $[n,k,d]$ optimal locally recoverable code with locality r and $n\ge\frac{q^{1+\frac{1}{t-1}}}{2t^2(r+1)^{1+\frac{2}{t-1}}}$. The parity check matrix of this code has the form of (6). Moreover, if d is a constant, there exists a deterministic algorithm running in polynomial time to construct this code.*

5 The connection with extremal graph theory

To our surprise, it turns out that finding a collection of sets satisfying (8) is equivalent to constructing an $(r+1)$-uniform hypergraph avoiding short Berge cycle. The latter is one of the central problems in extremal graph theory and this problem is extremely difficult.

▶ **Lemma 18.** *There exist m sets satisfying (8) if and only if there exists an $(r+1)$-hypergraph $H=([q],E)$ with $|E|=m$ that does not have any Berge ℓ-cycles for all $\ell\le t$.*

Proof. To see the equivalence of these two problems, we define an $(r+1)$-hypergraph as follows: Let $H=(V,E)$ with $V=[q]$ and $E=\{A_1,\dots,A_m\}$. It is clear that H is an $(r+1)$-hypergraph. Assume that there exists $k\le t$ subsets $A_{i_1},\dots,A_{i_k}$ does not satisfy the condition that $|\bigcup_{j=1}^{k}A_{i_j}|\ge rk+1$. The same argument in Theorem 12 implies that there exists a cycle $(1,2,\dots,\ell)$ such that $j\in A_{i_j}\cap A_{i_{j+1}}$. That means $\{j-1,j\}\subseteq A_{i_j}$ for $j=2,\dots,\ell$ and $\{1,\ell\}\subseteq A_{i_1}$. By the definition of Berge cycle, the $(r+1)$-hypergraph

H contains this Berge ℓ-cycle $(1, 2, \ldots, \ell)$. On the other hand, assume that there exists a Berge ℓ-cycle in H. Denote the ℓ edges of this cycle $A_{i_1}, \ldots, A_{i_\ell}$. The results follows since $|A_{i_j} \cap A_{i_{j+1}}| \geq 1$ for $i = 1, \ldots, \ell - 1$ and $|A_{i_1} \cap A_{i_\ell}| \geq 1$. ◀

The equivalence of both the problems allow us to make use of known results in this area. Let $\mathcal{F}$ be a family of $r+1$-graph. Denote by $ex_{r+1}(n, \mathcal{F})$ the maximum number of edges in an $(r+1)$-hypergraph that does not contain any subgraphs in $\mathcal{F}$. Denote by BC_k the set of k-cycles. Let $\mathcal{B}_k = \{BC_2, \ldots, BC_k\}$. One upper bound on $ex_{r+1}(n, \mathcal{B}_t)$ is obtained by reducing this problem to an $m \times n$ bipartite graph with girth more than $2t$ and apply the result in [7].

▶ **Proposition 19** ([19]). *$ex_{r+1}(n, \mathcal{B}_t)$ is upper bounded by*
- **(i)** $\frac{n}{r}(\frac{n}{r+1})^{\frac{2}{t-1}} + \frac{n}{r+1}$ *if t is odd,*
- **(ii)** $\frac{n}{r(r+1)} n^{\frac{2}{t}} + \frac{n}{r+1}$ *if t is even.*

Since these two problems are equivalent, Proposition 19 gives an upper bound on the number m of sets A_i. For $t = 3, 4$, the following two propositions show that this upper bound is asymptotically tight. However, constructing such hypergraph requires sophisticated knowledge in this area which is beyond the scope of this paper. We summarize the results as follows.

▶ **Proposition 20** ([18]). *There exists explicit construction of $(r+1)$-hypergraph $H = ([q], E)$ with $|E| = q^{2-o(1)}$ that contains no subgraph in $\mathcal{B}_3$.*

▶ **Proposition 21** ([17]). *There exists explicit construction of $(r+1)$-hypergraph $H = ([q], E)$ with $|E| = \Theta(q^{\frac{3}{2}-o(1)})$ that contains no subgraph in $\mathcal{B}_4$.*

Determining the exact value of $ex_{r+1}(n, \mathcal{B}_t)$ for $r \geq 2$ and $t \geq 3$ is extremely difficult. A major open problem in this area is whether $ex_{r+1}(n, \mathcal{B}_t) = \Omega(n^{1+\frac{2}{t}})$. A tighter lower bound for general t can be obtained from H-free random process [2]. The method in [2] can also be applied to hypergraph and add a log factor above the probabilistic method in Theorem 15. Again this technique is beyond our scope.

▶ **Proposition 22** ([18]). *$ex_{r+1}(n, \mathcal{B}_t) = \Omega_{r,t}(n(n \log n)^{\frac{1}{t-1}})$.*

Theorem 1 summarizes all above results in the language of codes.

6 An upper bound on the length of optimal LRC

In this section, we derive an upper bound on the length of optimal LRC. Our upper bound holds for all optimal LRC. The proof of our upper bound is a refined analysis of the argument of Theorem 3.3 in [6]. Recall the following Lemma in [6].

▶ **Lemma 23** (Lemma 3.1 [6]). *Let C be an $[n, k, d]_q$ linear optimal LRC with locality r. Then, there exist $\frac{n}{r+1}$ disjoint recovery sets, each of size $r+1$ provided that*

$$\frac{n}{r+1} \geq \left(d - 2 - \left\lfloor \frac{d-2}{r+1} \right\rfloor\right)(3r+2) + \left\lfloor \frac{d-2}{r+1} \right\rfloor + 1. \tag{9}$$

We use above lemma to force the optimal LRC to have disjoint recovery sets.

▶ **Theorem 24.** *Let C be an optimal $[n, k, d]_q$-linear locally repairable codes of locality r with $(r+1)|n$ and $n = \Omega(dr^2)$ satisfying (9) in Lemma 23. If $d \geqslant 5$, then*

$$n \leq \begin{cases} O(q^3) & \text{if } d \bmod r+1 > 5 \text{ or } d \bmod r+1 < 2, \\ O(q^2) & \text{if } 2 \leq d \bmod r+1 \leq 5. \end{cases} \tag{10}$$

Proof. Put $h = d - 2 - \lfloor \frac{d-2}{r+1} \rfloor$. Then $n - k = \frac{n}{r+1} + h$ and $h \le d - 2$. We first follow the same line of proof in Theorem 3.3, [6]. It allows us to assume that the parity-check matrix of optimal LRC C is as follows:

$$H = \left(\begin{array}{c|c|ccc|c} \mathbf{1} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{1} \\ \hdashline & & & A & & \end{array} \right), \tag{11}$$

where A is an $h \times n$ matrix over $\mathbb{F}_q$. The submatrix consisting of the first ℓ rows of H is a block diagonal matrix. Let $\mathbf{h}_{i,j}$ be the $((i-1)(r+1)+j)$-th column of H, i.e.,

$$\mathbf{h}_{i,j} = (\underbrace{0,\ldots,0}_{i-1}, 1, \underbrace{0,\ldots,0}_{\ell-i}, \mathbf{v}_{i,j})^T \tag{12}$$

for some $\mathbf{v}_{i,j} \in \mathbb{F}_q^h$, where T stands for transpose. Define

$$\mathbf{h}'_{i,j} := \mathbf{h}_{i,j} - \mathbf{h}_{i,r+1} = (\underbrace{0,\ldots,0}_{\ell}, \mathbf{v}_{i,j} - \mathbf{v}_{i,r+1})^T$$

for $i \in [\ell]$ and $j \in [r]$. It is clear that there are only $h = d - 2 - \lfloor \frac{d-2}{r+1} \rfloor$ nonzero components in $\mathbf{h}'_{i,j}$. Assume that $d - 6 = a(r+1) + b$ for some $0 \le b \le r$. Then, we claim that the first $ar + b$ columns $\mathbf{h}'_{i,j}$ are linearly independent. This is because the linear combination of these $ar + b$ columns leads to

$$\sum_{i=1}^{a} \sum_{j=1}^{r} \lambda_{i,j}(\mathbf{h}_{i,j} - \mathbf{h}_{i,r+1}) + \sum_{j=1}^{b} \lambda_{a+1,j}(\mathbf{h}_{a+1,j} - \mathbf{h}_{a+1,r+1}) \ne 0$$

as it is a linear combination of $ar + b + a + 1 = d - 5$ columns of H.

Since these $ar + b$ columns $\mathbf{h}'_{i,j}$ are linearly independent, there exist $ar + b$ indices where these $ar + b$ vectors $\mathbf{h}'_{i,j}$ span the whole space. Without loss of generality, we assume they are the last $ar + b$ indices. To simply our argument, we denote by S the index set of first $ar + b$ columns and $\bar{S}$ the index set of the rest columns. For each $(x, y) \in \bar{S}$, there exist $\lambda_{x_i,y_j} \in \mathbb{F}_q$ such that $\bar{\mathbf{h}}'_{x,y} = \mathbf{h}'_{x,y} - \sum_{(i,j)\in S} \lambda_{x_i,y_j} \mathbf{h}'_{i,j}$ gives a vector whose value on last $ar + b$ indices are all zero. The number of nonzero components of $\mathbf{h}'_{x,y}$ is at most

$$h - (ar + b) = d - 2 - \lfloor \frac{d-2}{r+1} \rfloor - (d - 6) + \lfloor \frac{d-6}{r+1} \rfloor = 4 + \lfloor \frac{d-6}{r+1} \rfloor - \lfloor \frac{d-2}{r+1} \rfloor.$$

On the other hand, let $\bar{\mathbf{h}}'_{x_1,y_1}$ and $\bar{\mathbf{h}}'_{x_2,y_2}$ be any two vectors of $\bar{\mathbf{h}}'_{x,y}, (x, y) \in \bar{S}$. Observe that they lie in the space spanned by $ar + b$ columns $\mathbf{h}'_{i,j}$ and $\mathbf{h}'_{x_1,y_1}, \mathbf{h}'_{x_2,y_2}$ which in turn contained in the space spanned by the first $d - 5$ columns $\mathbf{h}_{i,j}$ together with $\mathbf{h}_{x_1,y_1}, \mathbf{h}_{x_1,r+1}, \mathbf{h}_{x_2,y_2}, \mathbf{h}_{x_2,r+1}$. This implies that $\bar{\mathbf{h}}'_{x_1,y_1}$ and $\bar{\mathbf{h}}'_{x_2,y_2}$ are linearly independent. Let H_1 be the matrix whose columns consist of $\bar{\mathbf{h}}'_{x,y}, (x, y) \in \bar{S}$. Remove all zero rows in H_1 and denote the resulting matrix H_2. It is clear that any two columns of H_2 are linearly independent and H_2 has at most $4 + \lfloor \frac{d-6}{r+1} \rfloor - \lfloor \frac{d-2}{r+1} \rfloor$ rows. Let C_1 be a linear code whose parity-check matrix is H_2. We divide our discussion into two cases.

1. $\lfloor \frac{d-6}{r+1} \rfloor - \lfloor \frac{d-2}{r+1} \rfloor = 0$, i.e., $d \bmod r + 1 > 5$ or $d \bmod r + 1 < 2$.
Then, C_1 has length $N \ge \frac{rn}{r+1} - d$, dimension $k \ge N - 4$ and distance $d \ge 3$. In the worst scenario, we assume that $k = N - 4$ and $d = 3$. Applying the Hamming bound on code C_1 gives $q^{N-4} \le \frac{q^N}{N(q-1)}$. This implies $N \le q^3$, i.e., $n \le \frac{r+1}{r}(d + q^3)$.

2. $\lfloor\frac{d-6}{r+1}\rfloor - \lfloor\frac{d-2}{r+1}\rfloor = -1$ i.e., $2 \le d \bmod r+1 \le 5$.
Then, C_1 has length $N \ge \frac{rn}{r+1} - d$, dimension $k \ge N-3$ and distance $d \ge 3$. In the worst scenario, we assume that $k = N-3$ and $d = 3$. Applying the Hamming bound on code C_1 gives $q^{N-3} \le \frac{q^N}{N(q-1)}$. This implies $N \le q^2$, i.e., $n \le \frac{r+1}{r}(d+q^2)$.

The proof is completed. ◀

References

1. Alexander Barg, Kathryn Haymaker, Everett W. Howe, Gretchen L. Matthews, and Anthony Várilly-Alvarado. Locally recoverable codes from algebraic curves and surfaces. *CoRR*, abs/1701.05212, 2017. `arXiv:1701.05212`.
2. Tom Bohman and Peter Keevash. The early evolution of the H-free process. *Inventiones mathematicae*, 181(2):291–336, August 2010.
3. Belá Bollobás. *Modern Graph Theory*. Springer, New York, 1998.
4. Parikshit Gopalan, Cheng Huang, Huseyin Simitci, and Sergey Yekhanin. On the Locality of Codeword Symbols. *IEEE Trans. Information Theory*, 58(11):6925–6934, 2012.
5. Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced Expanders and Randomness Extractors from Parvaresh–Vardy Codes. *J. ACM*, 56(4):20:1–20:34, July 2009.
6. Venkatesan Guruswami, Chaoping Xing, and Chen Yuan. How Long Can Optimal Locally Repairable Codes Be? In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2018, August 20-22, 2018 - Princeton, NJ, USA*, pages 41:1–41:11, 2018.
7. Shlomo Hoory. The Size of Bipartite Graphs with a Given Girth. *J. Comb. Theory, Ser. B*, 86(2):215–220, 2002.
8. Cheng Huang, Minghua Chen, and Jin 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), 12 - 14 July 2007, Cambridge, MA, USA*, pages 79–86, 2007.
9. Lingfei Jin. Explicit construction of optimal locally recoverable codes of distance 5 and 6 via binary constant weight codes. *CoRR*, abs/1808.04558, 2018. `arXiv:1808.04558`.
10. Lingfei Jin, Liming Ma, and Chaoping Xing. Construction of optimal locally repairable codes via automorphism groups of rational function fields. *CoRR*, abs/1710.09638, 2017. `arXiv:1710.09638`.
11. Xudong Li, Liming Ma, and Chaoping Xing. Optimal locally repairable codes via elliptic curves. *CoRR*, abs/1712.03744, 2017. `arXiv:1712.03744`.
12. Yuan Luo, Chaoping Xing, and Chen Yuan. Optimal locally repairable codes of distance 3 and 4 via cyclic codes. *CoRR*, abs/1801.03623, 2018. `arXiv:1801.03623`.
13. N. Prakash, Govinda M. Kamath, V. Lalitha, and P. Vijay Kumar. Optimal linear codes with a local-error-correction property. In *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*, pages 2776–2780, 2012.
14. Natalia Silberstein, Ankit Singh Rawat, Onur Ozan Koyluoglu, and Sriram Vishwanath. Optimal locally repairable codes via rank-metric codes. In *Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7-12, 2013*, pages 1819–1823, 2013.
15. Itzhak Tamo and Alexander Barg. A Family of Optimal Locally Recoverable Codes. *IEEE Trans. Information Theory*, 60(8):4661–4676, 2014.
16. Itzhak Tamo, Dimitris S. Papailiopoulos, and Alexandros G. Dimakis. Optimal Locally Repairable Codes and Connections to Matroid Theory. *IEEE Trans. Information Theory*, 62(12):6661–6671, 2016.
17. Craig Timmons and Jacques Verstraëte. A counterexample to sparse removal. *European Journal of Combinatorics*, 44:77–86, 2015.
18. Jacques Verstraëte. Personal communication.
19. Jacques Verstraëte. *Extremal problems for cycles in graphs*, pages 83–116. Springer International Publishing, Cham, 2016.