On Optimal Locally Repairable Codes with Super-Linear Length

Han Cai*, Ying Miao†, Moshe Schwartz‡, and Xiaohu Tang§
*Electrical & Computer Engineering, Ben-Gurion University of the Negev, Israel, hancai@aliyun.com
†Engineering, Information and Systems, University of Tsukuba, Japan, miao@sk.tsukuba.ac.jp
‡Electrical & Computer Engineering, Ben-Gurion University of the Negev, Israel, schwartz@ee.bgu.ac.il
§Information Science and Technology, Southwest Jiaotong University, China, xhutang@swjtu.edu.cn

***Abstract*—Optimal locally repairable codes with respect to the bound presented by Prakash *et al.* are considered. New upper bounds on the length of such optimal codes are derived. The new bounds both improve and generalize previously known bounds. Optimal codes are constructed, whose length is order optimal when compared with the new upper bounds. The length of the codes is super linear in the alphabet size.**

I. Introduction

Locally repairable codes were introduced to improve the efficiency of the repair process of a failed node [11] for codes applied in distributed storage systems. More precisely, locally repairable codes ensure that a failed symbol can be recovered by accessing only $r \ll k$ other symbols [11].

However, the original concept of locality only works when exactly one erasure occurs (that is, one node fails). Over the past few years, several generalizations have been suggested for the definition of locality. As examples we mention locality with a single repair set tolerating multiple erasures [18], locality with disjoint multiple repairable sets [25], [20], [5], and hierarchical locality [21].

In this paper, we focus on locally repairable codes with a single repair set that can repair multiple erasures locally [18]. By ensuring $\delta - 1 \geqslant 2$ redundancies in each repair set, this kind of locally repairable codes guarantees the system can recover from $\delta - 1$ erasures by accessing r surviving code symbols for each erasure. This is denoted as (r, δ)-locality.

Research on codes with (r, δ)-locality has proceeded along two main tracks. In the first track, upper bounds on the minimum Hamming distance and the code length have been studied. Singleton-type bounds were introduced for codes with (r, δ)-locality in [18], [22], [26]. In [4], a bound depending on the size of the alphabet was derived for the Hamming distance of codes with (r, δ)-locality. Via linear programming, another bound related with the size of the alphabet was introduced in [1]. Very recently, in [10], an interesting connection between the length of optimal linear codes with $(r, \delta = 2)$-locality and the size of the alphabet was derived.

In the second research track, constructions for optimal locally repairable codes have been studied. In [19], a construction of optimal locally repairable codes was introduced based on Gabidulin codes. By analyzing the structure of repair sets, optimal locally repairable codes were also constructed in [22]. In [24], a construction of optimal locally repairable codes with $q = \Theta(n)$ was proposed. In [23] and [27], optimal locally repairable codes were constructed using matroid theory. The construction of [24] was generalized in [15] to include more flexible parameters when $n \leqslant q$. Very recently, in [17], cyclic optimal locally repairable codes with unbounded length were constructed for Hamming distance $d = 3, 4$. Finally, for the case of Hamming distance $d = 5$, [10], [12], [3] presented constructions of locally repairable codes that have optimal distance as well as order-optimal length $n = \Theta(q^2)$.

In this paper we first prove that the bound in [10] holds for some other cases besides the one mentioned in [10]. We then derive a new upper bound on the length of optimal locally repairable codes for the case of $\delta > 2$. Finally, we give a general construction of locally repairable codes with length that is super-linear in the field size. Based on some special structures such as packings and Steiner systems, locally repairable codes are obtained with optimal Hamming distances and order-optimal length $\Omega(q^\delta)$ when $\delta > 2$. Thus, the bound for $\delta > 2$ is also asymptotically tight for some special cases.

The remainder of this paper is organized as follows. Section II introduces some preliminaries about locally repairable codes. Section III establishes an upper bound on the length of optimal locally repairable codes for the case $\delta > 2$. Section IV presents a construction of optimal locally repairable codes with length $n > q$. Section V concludes this paper with some remarks. For lack of space we omit all proofs, which are available in a full version of this work [6].

II. Preliminaries

We present the notation and basic definitions used throughout the paper. For a positive integer $n \in \mathbb{N}$, we define $[n] = \{1, 2, \ldots, n\}$. For any prime power q, let $\mathbb{F}_q$ denote the finite field with q elements. An $[n, k]_q$ linear code $\mathcal{C}$ over $\mathbb{F}_q$ is a k-dimensional subspace of $\mathbb{F}_q^n$ with a $k \times n$ generator matrix $G = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n)$, where $\mathbf{g}_i$ is a column vector of dimension k for all $i \in [n]$. Specifically, it is called an $[n, k, d]_q$ linear code if the minimum Hamming distance is d. For a subset $S \subseteq [n]$, let $|S|$ denote the cardinality of S, let 2^S denote the set of all subsets of S, and define

$$\mathrm{Rank}(S) = \mathrm{Rank}(\mathrm{Span}\,\{\mathbf{g}_i | i \in S\}).$$

H. Cai and M. Schwartz were supported in part by a German Israeli Project Cooperation (DIP) grant under grant no. PE2398/1-1.

X. Tang was supported in part by by the National Natural Science Foundation of China under Grant 61871331.

Y. Miao was supported by JSPS Grant-in-Aid for Scientific Research (B) under Grant No. 18H01133.

In [8], Gopalan *et al.* introduce the following definition for the locality of code symbols. The ith ($1 \leqslant i \leqslant n$) code symbol c_i of an $[n,k,d]_q$ linear code $\mathcal{C}$ is said to have locality r ($1 \leqslant r \leqslant k$), if it can be recovered by accessing at most r other symbols in $\mathcal{C}$. More precisely, symbol locality can also be rigorously defined as follows.

Definition 1 ([8])**:** For any column $\mathbf{g}_i$ of G with $i \in [n]$, define $\mathrm{Loc}(\mathbf{g}_i)$ as the smallest integer r such that there exists an $(r+1)$-subset $R_i = \{i, i_1, i_2, \ldots, i_r\} \subseteq [n]$ satisfying

$$\mathbf{g}_i \in \mathrm{Span}(R_i \setminus \{i\}), \text{ i.e., } \mathbf{g}_i = \sum_{t=1}^{r} \lambda_t \mathbf{g}_{i_t} \quad \lambda_t \in \mathbb{F}_q. \quad (1)$$

Equivalently, for any codeword $C = (c_1, c_2, \ldots, c_n) \in \mathcal{C}$, the ith component

$$c_i = \sum_{t=1}^{r} \lambda_t c_{i_t} \quad \lambda_t \in \mathbb{F}_q.$$

Define $\mathrm{Loc}(S) = \max_{i \in S} \mathrm{Loc}(\mathbf{g}_i)$ for any set $S \subseteq [n]$. Then, an $[n,k,d]_q$ linear code $\mathcal{C}$ is said to have information locality r if there exists $S \subseteq [n]$ with $\mathrm{Rank}(S) = k$ satisfying $\mathrm{Loc}(S) = r$. Furthermore, an $[n,k,d]_q$ linear code $\mathcal{C}$ is said to have all symbol locality r if $\mathrm{Loc}([n]) = r$.

To guarantee that the system can locally recover from multiple erasures, say, $\delta - 1$ erasures, the definition of locality was generalized in [18] as follows.

Definition 2 ([18])**:** The jth column $\mathbf{g}_j$, $1 \leqslant j \leqslant n$, of a generator matrix G of an $[n,k]_q$ linear code $\mathcal{C}$ is said to have (r,δ)-locality if there exists a subset $S_j \subseteq [n]$ such that:

- $j \in S_j$ and $|S_j| \leqslant r + \delta - 1$; and
- the minimum Hamming distance of the punctured code $\mathcal{C}|_{S_j}$ obtained by deleting the code symbols c_t ($t \in [n] \setminus S_j$) is at least δ,

where the set S_j is also called an (r,δ) repair set of $\mathbf{g}_j$. The code $\mathcal{C}$ is said to have information (r,δ)-locality if there exists $S \subseteq [n]$ with $\mathrm{Rank}(S) = k$ such that for each $j \in S$, $\mathbf{g}_j$ has (r,δ)-locality. Furthermore, the code $\mathcal{C}$ is said to have all symbol (r,δ)-locality if all the code symbols have (r,δ)-locality.

In [18] (for the case $\delta = 2$ [8]), the following upper bound on the minimum Hamming distance of linear codes with information (r,δ)-locality was derived.

Lemma 1 ([18])**:** For an $[n,k,d]_q$ linear code with information (r,δ)-locality,

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

Additionally, a locally repairable code is said to be *optimal* if its minimum Hamming distance attains this bound with equality.

III. Bounds on the Length of Locally Repairable Codes

The goal of this section is to derive upper bounds on the length of optimal locally repairable codes. Throughout this section, let

$$n = (r + \delta - 1)w + m, \qquad k = ru + v,$$

where $\delta \geqslant 2$, $0 \leqslant m \leqslant r + \delta - 2$, and $0 \leqslant v \leqslant r - 1$ are all integers.

We first characterize the basic structure of (r,δ) repair sets for an optimal locally repairable codes with all symbol (r,δ)-locality under the conditions that $(r+\delta-1)|n$ and $v = 0$ or $u \geqslant 2(r - v + 1)$. The main method is to analyze:

1) The effect of the intersection between repair sets on the Hamming distance of the locally repairable codes;
2) The influence of short repair sets with length strictly less than $r + \delta - 1$.

As a result, under the conditions that $(r+\delta-1)|n$ and $v = 0$ or $u \geqslant 2(r - v + 1)$, we can prove that the bound in Lemma 1 is achievable only in the case that there exist some repair sets with size $r + \delta - 1$ that form a partition of $[n]$.

Theorem 1: Let $\mathcal{C}$ be an optimal $[n,k,d]_q$ linear code with all symbol (r,δ)-locality, where optimality is with respect to the bound in Lemma 1. Let $\Gamma \subseteq 2^{[n]}$ be the set of all possible (r,δ) repair sets. Write $k = ru + v$, for integers u and v, and $0 \leqslant v \leqslant r - 1$. If $(r+\delta-1)|n$, $k > r$, and additionally, if $u \geqslant 2(r - v + 1)$ or $v = 0$, then there exists a set of (r,δ) repair sets $\mathcal{S} \subseteq \Gamma$, such that all $R \in \mathcal{S}$ are of cardinality $|R| = r + \delta - 1$, and $\mathcal{S}$ is a partition of $[n]$.

Before going on to the main result, we briefly consider the special case of $\delta = 2$. This special case was studied in [10] and an upper bound on the length of optimal codes was established. While we obtain the exact same bound as [10], our bound is an improvement since it has more relaxed conditions. In particular, the bound of [10] requires $k = \Omega(dr^2)$ whereas we require $k = \Omega(r^2)$. We now provide the exact claim:

Corollary 1: Let $\mathcal{C}$ be an optimal $[n,k,d]_q$ code with all symbol locality r. If $d \geqslant 5$, $k > r$, $(r+1)|n$, and additionally, $r|k$ or $k \geqslant 2r^2 + 2r - (2r-1)\langle k \rangle_r$, then

$$n = \begin{cases} O\left(dq^{\frac{4(d-2)}{d-a}-1}\right), & \text{if } a = 1, 2, \\ O\left(dq^{\frac{4(d-3)}{d-a}-1}\right), & \text{if } a = 3, 4, \end{cases} \quad (3)$$

where $a \equiv d \pmod 4$ and $\langle k \rangle_r$ denotes the least nonnegative integer congruent to k modulo r.

We introduce another corollary that stems from Theorem 1. It slightly extends [22, Theorem 9], originally proved only for $r|k$, and has a very similar proof.

Corollary 2: Let $\mathcal{C}$ be an optimal $[n,k,d]_q$ linear code with all symbol (r,δ)-locality, where optimality is with respect to the bound in Lemma 1. If $k > r$, $n = w(r+\delta-1)$, and additionally $r|k$ or $k \geqslant 2r^2 + 2r - (2r-1)\langle k \rangle_r$, then there are w pairwise-disjoint (r,δ) repair sets, $R_1, \ldots, R_w \subseteq [n]$, such that for all $1 \leqslant i \leqslant w$, $|R_i| = r + \delta - 1$, and the punctured code $\mathcal{C}|_{R_i}$ is a linear $[r+\delta-1, r, \delta]_q$ MDS code.

We now extend our scope and consider locally repairable codes for the case of $\delta > 2$. To this end, we reduce locally repairable codes from the case $\delta > 2$ into $\delta = 2$.

Lemma 2: Let $n = w(r+\delta-1)$, $\delta > 2$, $k = ur + v > r$, and additionally, $u \geqslant 2(r-v+1)$ or $v = 0$, where all parameters are integers. If there exists an optimal $[n,k,d]_q$ linear code $\mathcal{C}$ with all symbol (r,δ)-locality, then there exists a $[w(r+1),k,d']_q$ linear code $\mathcal{C}'$ with all symbol $(r,2)$-locality (i.e., locality r), and $d' \geqslant 2\lfloor(d-1)/\delta\rfloor + 1$.

Now based on Lemma 2, the case $\delta > 2$ is closely related to the case $\delta = 2$. A similar proof of [10, Theorem 3.2] can help us to prove the following bound for $\delta > 2$.

Theorem 2: Let $n = w(r+\delta-1)$, $\delta > 2$, $k = ur + v$, and additionally, $u \geqslant 2(r-v+1)$ or $v = 0$, where all parameters are integers. Assume there exists an optimal $[n,k,d]_q$ linear code $\mathcal{C}$ with all symbol (r,δ)-locality, and define $t = \lfloor(d-1)/\delta\rfloor$. If $2t+1 > 4$, then

$$\begin{aligned} n &\leqslant \begin{cases} \frac{(t-1)(r+\delta-1)}{2r(q-1)} q^{\frac{2(w-u)r-2v}{t-1}}, & \text{if } t \text{ is odd} \\ \frac{t(r+\delta-1)}{2r(q-1)} q^{\frac{2(w-u)r-2v}{t}}, & \text{if } t \text{ is even} \end{cases} \\ &= O\left(\frac{t(r+\delta)}{r} q^{\frac{(w-u)r-v}{\lfloor t/2 \rfloor}-1}\right), \end{aligned}$$

where $w-u$ can also be rewritten as $w-u = \lfloor(d-1+v)/(r+\delta-1)\rfloor$.

For the case $d > r+\delta$, we can improve the bounds in Corollary 1 and Theorem 2 as follows.

Corollary 3: Let $n = w(r+\delta-1)$, $\delta > 2$, $k = ur + v > r$, and additionally, $u \geqslant 2(r-v+1)$ or $v = 0$, where all parameters are integers. If there exists an optimal $[n,k,d]_q$ linear code $\mathcal{C}$ with $d > r+\delta$ and all symbol (r,δ)-locality, then for $\delta = 2$

$$n \leqslant \epsilon(r+\delta-1) + \begin{cases} \frac{(d'-a)(r+1)}{4(q-1)r} q^{\frac{4(d'-2)}{d'-a}}, & \text{if } a = 1,2, \\ \frac{r+1}{r}\left(\frac{d'-a}{4(q-1)} q^{\frac{4(d'-3)}{d'-a}} + 1\right), & \text{if } a = 3,4, \end{cases}$$

and for $\delta > 2$

$$n \leqslant \epsilon(r+\delta-1) + \begin{cases} \frac{(t-1)(r+\delta-1)}{2r(q-1)} q^{\frac{2(w'-u)r-2v}{t-1}}, & \text{if } t \text{ is odd,} \\ \frac{t(r+\delta-1)}{2r(q-1)} q^{\frac{2(w'-u)r-2v}{t}}, & \text{if } t \text{ is even,} \end{cases}$$

where $\epsilon = \lceil(d-1)/(r+\delta-1)\rceil - 1$, $d' = d - \epsilon(r+\delta-1)$, $w' = w - \epsilon$, $a \equiv d' \pmod 4$, and $t = \lfloor(d'-1)/(\delta)\rfloor$. Herein, we assume that $2t+1 > 4$ holds.

IV. Optimal Locally Repairable Codes with Super-Linear Length

In this section, we introduce a construction of optimal locally repairable codes with length n that is super linear in the field size q. To streamline the presentation in the section we adopt a slightly different notation than the previous one: we use $n = w(r+\delta-1)$ and $k = (w-1)r+v$ for $0 < v \leqslant r$, where all parameters are integers.

Construction A: Let the k information symbols be partitioned into w sets, say,

$$\begin{aligned} I^{(i)} &= \{I_{(i,1)}, I_{(i,2)}, \ldots, I_{(i,r)}\}, \quad \text{for } i \in [w-1], \\ I^{(w)} &= \{I_{(w,1)}, I_{(w,2)}, \ldots, I_{(w,v)}\}. \end{aligned}$$

A linear code with length n is constructed by describing a linear map from the information $\boldsymbol{I} = (I_{(1,1)}, \ldots, I_{(w,v)}) \in \mathbb{F}_q^k$ to a codeword $\boldsymbol{C}(\boldsymbol{I}) = (c_{1,1}, \ldots, c_{w,r+\delta-1}) \in \mathbb{F}_q^n$, thus the $[n,k]_q$ linear code is $\mathcal{C} = \{\boldsymbol{C}(\boldsymbol{I}) : \boldsymbol{I} \in \mathbb{F}_q^k\}$. This mapping is performed by the following three steps:

a) Step 1 – Partial parity check symbols: For $1 \leqslant i \leqslant w-1$, let $S_i = \{\theta_{i,t} : 1 \leqslant t \leqslant r+\delta-1\} \subseteq \mathbb{F}_q$ and let $f_i(x)$ be the unique polynomial over $\mathbb{F}_q$ with $\deg(f_i) \leqslant r-1$ that satisfies $f_i(\theta_{i,t}) = I_{i,t}$ for $1 \leqslant t \leqslant r$. For $1 \leqslant i \leqslant w-1$ and $1 \leqslant t \leqslant r+\delta-1$, set $c_{i,t} = f_i(\theta_{i,t})$.

b) Step 2 – Auxiliary symbols: Let $\{\alpha_t : 1 \leqslant t \leqslant r-v\} \subseteq \mathbb{F}_q \setminus (\bigcup_{1 \leqslant i \leqslant w-1} S_i)$. For $1 \leqslant i \leqslant w-1$, and $1 \leqslant t \leqslant r-v$, define

$$a_{i,t} = \frac{f_i(\alpha_t)}{\prod_{\theta \in S_i}(\alpha_t - \theta)}. \tag{4}$$

c) Step 3 – Global parity check symbols: Let $S_w = \{\theta_{w,t} : 1 \leqslant t \leqslant r+\delta-1\} \subseteq \mathbb{F}_q \setminus \{\alpha_t : 1 \leqslant t \leqslant r-v\}$ and let $f_w(x)$ be the unique polynomial over $\mathbb{F}_q$ with $\deg(f_w) \leqslant r-1$ that satisfies $f_w(\theta_{w,t}) = I_{w,t}$ for $1 \leqslant t \leqslant v$, as well as

$$\sum_{1 \leqslant i \leqslant w} a_{i,t} = 0 \text{ for } 1 \leqslant t \leqslant r-v, \tag{5}$$

where $a_{w,t} = \frac{f_w(\alpha_t)}{\prod_{\theta \in S_w}(\alpha_t - \theta)}$ for $1 \leqslant t \leqslant r-v$. Here, the polynomial $f_w(x)$ can be viewed as a polynomial over $\mathbb{F}_q$ determined by $I_{(w,j)}$, $1 \leqslant j \leqslant v$ and $a_{w,t}$ for $1 \leqslant t \leqslant r-v$. Thus, $f_w(x)$ is unique and well defined. Set $c_{w,j} = f_w(\theta_{w,j})$, for $1 \leqslant j \leqslant r+\delta-1$.

Remark 1: At first glance there appears to be a distinction between code symbols $c_{i,j}$ with $1 \leqslant i \leqslant w-1$ and those with $i = w$. However, careful thought reveals that the code symbols that correspond to the sets S_i for $1 \leqslant i \leqslant w$ are essentially symmetric, i.e., any $w-1$ sets of code symbols can determine v code symbols of the remaining set according to (5).

The all symbol (r,δ)-locality of the code $\mathcal{C}$ generated by Construction A directly follows from Steps 1 and 3. The problem that remains is to determine the minimum Hamming distance of $\mathcal{C}$. By restricting the structure of the evaluation points, i.e., $S_1, S_2, \cdots, S_w$, the Hamming distance can be lower bounded as follows.

Theorem 3: Let μ be a positive integer, and let S_i, $i \in [w]$ be the sets defined in Construction A. If every subset $\mathcal{R} \subseteq \{S_i : 1 \leqslant i \leqslant w\}$, $|\mathcal{R}| = \mu$, satisfies that for all $S' \in \mathcal{R}$,

$$\left| S' \cap \left(\bigcup_{S \in \mathcal{R} \setminus \{S'\}} S \right) \right| < \delta, \tag{6}$$

then the code $\mathcal{C}$ generated by Construction A is an $[n,k,d]_q$ linear code, with $d \geqslant \min\{r-v+\delta, (\mu+1)\delta\}$ and all symbol (r,δ)-locality, where $n = w(r+\delta-1)$, $k = (w-1)r+v$, $1 \leqslant v \leqslant r$, and all parameters are integers.

According to Lemma 1, the Hamming distance of $\mathcal{C}$ is upper bounded by $r-v+\delta$. Therefore, the key in applying Theorem 3 to find optimal locally repairable codes is to find sets $S_1, \ldots, S_w$ of evaluation points with $\mu\delta \geqslant r-v$. In the meantime, we also need to find as many evaluation points as possible to get a long code. We first mention trivial such families of sets that allow optimal codes with length $n > q$.

Corollary 4: Let $n = w(r+\delta-1)$, $k = (w-1)r+v$, $1 \leqslant v \leqslant r$, be integers. If $r-v \leqslant \delta$ and $q \geqslant 2r+\delta-1-v$ then there exists an optimal $[n,k,d = r-v+\delta]_q$ linear code with all symbol (r,δ)-locality, where optimality is with respect to the bound in Lemma 1.

Remark 2: We remark that in the case described in Corollary 4, we can let $S_i = S_j$ for $1 \leqslant i \neq j \leqslant w$. Thus, the length of the code $\mathcal{C}$ can be as long as we wish. This result is already known for the case $\delta = 2$ and $d \leqslant 4$ (see [17]), and is, to the best of our knowledge, new for the case of $\delta > 2$. This result also shows that the condition $2t+1 > 4$ is necessary for Theorem 2, since the code length is unbounded for the case $2t+1 \leqslant 4$, i.e., $t \leqslant 1$ corresponding to the case $r-v \leqslant \delta$, where $t = \lfloor (d-1)/\delta \rfloor = \lfloor \frac{r+v+\delta-1}{\delta} \rfloor$.

Corollary 5: Let $n = w(r+\delta-1)$, $k = (w-1)r+v$, $1 \leqslant v \leqslant r$, be integers. Let $S \subseteq \mathbb{F}_q \setminus \{\alpha_i \,:\, 1 \leqslant i \leqslant r-v\}$, $|S| = \delta-1$, be a fixed subset. Take $S_i \subseteq \mathbb{F}_q \setminus \{\alpha_i \,:\, 1 \leqslant i \leqslant r-v\}$ for $1 \leqslant i \leqslant w$. If $S_i \cap S_j \subseteq S$ for $1 \leqslant i \neq j \leqslant w$, then the code $\mathcal{C}$ generated by Construction A is an optimal $[n,k,d = r-v+\delta]_q$ linear code with all symbol (r,δ)-locality, where optimality is with respect to the bound in Lemma 1.

Remark 3: By Corollary 5, Construction A can generate optimal linear codes with all symbol (r,δ)-locality with flexible parameters and with $n > q$, since we can simply select S as a common part of S_i for $1 \leqslant i \leqslant w$. In [15], optimal locally repairable codes are also constructed with flexible parameters. However, in [15] the construction is based on the so-called good polynomials [24], [16] and $n \leqslant q$.

A combinatorial structure that captures the interaction between the evaluation-point sets, $S_1, \ldots, S_w$, in Construction A is a union-intersection-bounded family [9]. Its definition is now given:

Definition 3 ([9])**:** Let n_1, τ, δ, t, s be positive integers such that $n_1 \geqslant \tau \geqslant 2$, $\tau \geqslant \delta$ and $t \geqslant s$. The $(s,t;\delta)$-union-intersection-bounded family (denoted by $(s,t;\delta)$-UIBF(τ, n_1)) is a pair $(\mathcal{X}, \mathcal{S})$, where $\mathcal{X}$ is a set of n_1 elements (called points) and $\mathcal{S} \subseteq 2^{\mathcal{X}}$ is a collection of τ-subsets of $\mathcal{X}$ (called blocks), such that any $s+t$ distinct blocks $A_1, A_2, \ldots, A_s, B_1, B_2, \ldots, B_t \in \mathcal{S}$ satisfy

$$\left| \left(\bigcup_{1 \leqslant i \leqslant s} A_i \right) \cap \left(\bigcup_{1 \leqslant i \leqslant t} B_i \right) \right| < \delta.$$

The following corollary follows from Theorem 3 and Lemma 1.

Corollary 6: Let $n = w(r+\delta-1)$, $k = (w-1)r+v$, $1 \leqslant v \leqslant r$, be integers, and let μ be a positive integer with $\mu\delta \geqslant r-v$. If $(\mathbb{F}_q \setminus \{\alpha_t \,:\, 1 \leqslant t \leqslant r-v\}, \mathcal{S} = \{S_i \,:\, 1 \leqslant i \leqslant w\})$ is a $(1,\mu-1;\delta)$-UIBF$(r+\delta-1, q-r+v)$ then the code $\mathcal{C}$ generated by Construction A is an optimal $[n,k,d = r-v+\delta]_q$ linear code with all symbol (r,δ)-locality, where optimality is with respect to the bound in Lemma 1.

In [9], a lower bound on the size of $(1,\mu-1;\delta)$-UIBF$(r+\delta-1, q)$ is given, which immediately implies a lower bound on the length of the codes generated by Construction A according to Corollary 6.

Lemma 3 ([9])**:** Let μ, δ, r be positive integers, then there exists a $(1,\mu-1;\delta)$-UIBF$(r+\delta-1, q)$ $(\mathbb{F}_q, \mathcal{S})$ with $|\mathcal{S}| = \Omega(q^{\frac{\delta}{\mu}})$, where r, δ, μ are regarded as constants.

Based on Corollary 6 and Lemma 3, we have the following:

Corollary 7: Let $n = w(r+\delta-1)$, $k = (w-1)r+v$, $1 \leqslant v \leqslant r$, be integers, and let μ be a positive integer with $\mu\delta \geqslant r-v$. Then Construction A can generate an optimal (with respect to the bound in Lemma 1) $[n,k,d = r-v+\delta]_q$ linear code $\mathcal{C}$ with all symbol (r,δ)-locality and length $n = O(q^{\frac{\delta}{\mu}})$, where we regard r, δ, and μ as constants.

In what follows, we consider some special sufficient conditions for (6) to construct optimal linear codes with all symbol (r,δ)-locality.

Theorem 4: Let $n = w(r+\delta-1)$, $k = (w-1)r+v$, $1 \leqslant v \leqslant r$, be integers, and let a be a positive integer. If $|S_i \cap S_j| \leqslant a$ for $1 \leqslant i \neq j \leqslant w$ and $r-v < \frac{\delta^2}{a}$, then the code $\mathcal{C}$ generated by Construction A is an optimal $[n,k,d = r-v+\delta]_q$ linear code with all symbol (r,δ)-locality, where optimality is with respect to the bound in Lemma 1.

Definition 4 ([7], VI. 40)**:** Let $n_1 \geqslant 2$ be an integer and u a positive integer. A τ-$(n_1,t,1)$-*packing* is a pair $(\mathcal{X}, \mathcal{S})$, where $\mathcal{X}$ is a set of n_1 elements (called points) and $\mathcal{S} \subseteq 2^{\mathcal{X}}$ is a collection of t-subsets of $\mathcal{X}$ (called blocks), such that each τ-subset of $\mathcal{X}$ is contained in at most one block of $\mathcal{S}$. Furthermore, if each τ-subset of $\mathcal{X}$ is contained in exactly one block of $\mathcal{S}$, then $(\mathcal{X}, \mathcal{S})$ is also called a (τ, t, n_1)-*Steiner system*.

The following corollary follows directly from Theorem 4.

Corollary 8: Let $n_1 = q-r+v$. If there exists a $(\tau+1)$-$(n_1, r+\delta-1, 1)$-packing with blocks $\mathcal{S}$ and $r-v < \frac{\delta^2}{\tau}$, then there exists an optimal $[n,k,d]_q$ linear code with all symbol (r,δ)-locality, where $n = |\mathcal{S}|(r+\delta-1)$, $k = (|\mathcal{S}|-1)r+v$, and $d = r-v+\delta$.

The number of blocks of a packing is upper bound by the following Johnson bound [13]:

Lemma 4: ([13]) The maximum possible number of blocks of a $(\tau+1)$-$(n_1, r+\delta-1, 1)$-packing $\mathcal{S}$ satisfies

$$|\mathcal{S}| \leqslant \left\lfloor \frac{n_1}{r+\delta-1} \left\lfloor \frac{n_1-1}{r+\delta-2} \cdots \left\lfloor \frac{n_1-\tau}{r+\delta-1-\lambda} \right\rfloor \cdots \right\rfloor \right\rfloor.$$

Thus, the number of blocks for a $(\tau+1)$-$(n_1, r+\delta-1, 1)$-packing can be as large as $O(n_1^{\tau+1})$, when τ, r, and δ are regarded as constants.

Corollary 9: Let $n_1 = q-r+v$. If there exists a $(\tau+1)$-$(n_1, r+\delta-1, 1)$-packing with blocks $\mathcal{S}$, $|\mathcal{S}| = O(n_1^{\tau+1})$, and $r-v < \frac{\delta^2}{\tau}$, then there exists an optimal $[n,k,d]_q$ linear code with all symbol (r,δ)-locality, where $n = |\mathcal{S}|(r+\delta-1) = O(q^{\tau+1})$, $k = (|\mathcal{S}|-1)r+v$ and $d = r-v+\delta$. In particular, for the case $w-1 \geqslant 2(r-v+1)$, $r-v = \delta+1$, i.e., $d = 2\delta+1$

and $\tau = \delta - 1$, the code based on the $(\tau+1)$-$(n_1, r+\delta-1, 1)$-packing has asymptotically optimal length, where r and δ are regarded as constants.

As an example, we also analyze the length of the codes based on Steiner systems.

Corollary 10: Let $n_1 = q - r + v$. If there exists a $(\tau + 1, r + \delta - 1, n_1)$-Steiner system and $r - v \leqslant \frac{\delta^2}{\tau}$, then there exists an optimal $[n, k, d]_q$ linear code with all symbol (r, δ)-locality, where

$$n = \frac{\binom{n_1}{\tau+1}(r+\delta-1)}{\binom{r+\delta-1}{\tau+1}}, \quad k = \left(\frac{\binom{n_1}{\tau+1}}{\binom{r+\delta-1}{\tau+1}} - 1\right) r + v,$$

and $d = r-v+\delta$. In particular, for the case $w-1 \geqslant 2(r-v+1)$, $r-v = \delta+1$, i.e., $d = 2\delta+1$ and $\tau = \delta-1$, the code based on the $(\delta, r+\delta-1, q-\delta-1)$-Steiner system has asymptotically optimal length, where r and δ are regarded as constants.

Remark 4: For the case $\delta = 2$ and $d = 5$, optimal linear codes with all symbol $(r, 2)$-locality and asymptotically optimal length $\Theta(q^2)$ have been introduced in [10], [12].

Remark 5: Given positive integers τ, r and $\delta > 2$, the natural necessary conditions for the existence of a $(\tau+1, r+\delta-1, q-r+v)$-Steiner system are that $\binom{q-r+v-i}{\tau+1-i} | \binom{r+\delta-1-i}{\tau+1-i}$ for all $0 \leqslant i \leqslant \tau$. It was shown in [14] that these conditions are also sufficient except perhaps for finitely many cases. While q might not be a prime power, any prime power $\overline{q} \geqslant q$ will suffice for our needs. It is known, for example, that there is always a prime in the interval $[q, q + q^{21/40}]$ (see [2]). Thus, Construction A provides infinitely many optimal linear $[n, k, d]_{\overline{q}}$ codes, with all symbol (r, δ)-locality, and

$$\begin{aligned}
&n = (r+\delta-1) \cdot \frac{\binom{q-r+v}{\tau+1}}{\binom{r+\delta-1}{\tau+1}} = \Omega(q^{\tau+1}) = \Omega(\overline{q}^{\tau+1}), \\
&k = \left(\frac{\binom{q-r+v}{\tau+1}}{\binom{r+\delta-1}{\tau+1}} - 1\right) r + v, \\
&d = r - v + \delta,
\end{aligned}$$

i.e., with length super-linear in the field size.

V. Concluding Remarks

In this paper, we first derived an upper bound for the length of optimal locally repairable codes when $\delta > 2$. As a byproduct, we also extended the range of parameters for the known bound (the case $\delta = 2$) and improve its performance for the case $d > r + \delta$. A general construction of locally repairable codes was introduced. By the construction, locally repairable codes with length super-linear in the field size can be generated. In particular, for some cases those codes have asymptotically optimal length with respect to the new bound.

Several combinatorial structures, e.g., union-intersection-bounded families, packings, and Steiner systems, satisfy (6) and play a key role in determining the length of the codes generated by Construction A. If more of those structures with a large number of blocks can be constructed, more good codes with length $n > q$ can be generated. Finding more such combinatorial structures and explicit constructions for them, is left for future research.

References

[1] A. Agarwal, A. Barg, S. Hu, A. Mazumdar, and I. Tamo, "Combinatorial alphabet-dependent bounds for locally recoverable codes," *IEEE Trans. Inf. Theory,* vol. 64, no. 5, pp. 3481-3492, 2018.

[2] R.C. Baker, G. Harman, and J. Pintz, "The difference between consecutive primes, II," *Proceedings of the London Mathematical Society*, vol. 83, no. 3, pp. 532–562, Nov. 2001.

[3] A. Beemer, R. Coatney, V. Guruswami, H.H. López, and F. Piñero, "Explicit optimal-length locally repairable codes of distance 5," *arXiv:1810.03980*, 2018.

[4] V.R. Cadambe and A. Mazumdar, "Bounds on the size of locally recoverable codes," *IEEE Trans. Inf. Theory,* vol. 61, no. 11, pp. 5787-5794, Nov. 2015.

[5] H. Cai, Y. Miao, M. Schwartz, and X. Tang, "On optimal locally repairable codes with multiple disjoint repair sets," submitted to *IEEE Trans. Inf. Theory*.

[6] H. Cai, Y. Miao, M. Schwartz, and X. Tang, "On optimal locally repairable codes with super-linear length," *arXiv:1812.11942*, 2018.

[7] C. Colbourn and J. Dinitz, *Handbook of Combinatorial Designs, Second Edition,* Chapman & Hall/CRC, 2006.

[8] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *IEEE Trans. Inf. Theory,* vol. 58, no. 11, pp. 6925-6934, Nov. 2012.

[9] Y. Gu and Y. Miao "Union-intersection-bounded families and their applications," to appear in *Discrete Applied Mathematics*.

[10] V. Guruswami, C. Xing, and C. Yuan, "How long can optimal locally repairable codes be?" *arXiv:1807.01064*, 2018.

[11] C. Huang, M. Chen, and J. Li, "Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems," In NCA, 2007.

[12] L. Jin, "Explicit construction of optimal locally recoverable codes of distance 5 and 6 via binary constant weight codes," *arXiv:1808.04558*, 2018.

[13] S.M. Johnson, "A new upper bound for error-correcting codes," *IEEE Trans. Inf. Theory,* vol. 8, vol. 2, pp. 203-207, Apr. 1962.

[14] P. Keevash, "The existence of designs," *arXiv:1401.3665v1*, 2014.

[15] O. Kolosov, A. Barg, I. Tamo, and G. Yadgar, "Optimal LRC codes for all lenghts $n \leqslant q$," *arXiv:1802.00157*, 2018.

[16] J. Liu , S. Mesnager, and L. Chen, "New constructions of optimal locally recoverable codes via good polynomials," *IEEE Trans. Inf. Theory,* vol. 64, no. 2, pp. 889-899, Feb. 2018.

[17] Y. Luo, C. Xing, and C. Yuan, "Optimal locally repairable codes of distance 3 and 4 via cyclic codes," *IEEE Trans. Inf. Theory,* vol. 65, no. 2, pp. 1048-1053, Feb. 2019.

[18] N. Prakash, G.M. Kamath, V. Lalitha, and P.V. Kumar, "Optimal linear codes with a local-error-correction property," In Proc. of IEEE ISIT, 2012.

[19] A.S. Rawat, O.O. Koyluoglu, N. Silberstein, and S. Vishwanath, "Optimal locally repairable and secure codes for distributed storage systems," *IEEE Trans. Inf. Theory,* vol. 60, no. 1, pp. 212-236, Jan. 2014.

[20] A.S. Rawat, D.S. Papailopoulos, A.G. Dimakis, and S. Vishwanath, "Locality and availability in distributed storage," *IEEE Trans. Inf. Theory,* vol. 62, no. 8, pp. 4481-4493, Aug. 2016.

[21] B. Sasidharan, G.K. Agarwal, and P.V. Kumar, "Codes with hierarchical locality," In Proc. of IEEE ISIT, 2015.

[22] W. Song, S.H. Dau, C. Yuen, and T.J. Li, "Optimal locally repairable linear codes," *IEEE J. Slect. Areas Commun.,* vol. 32, no. 5, pp. 1019-1036, May 2014.

[23] I. Tamo, D. Papailiopoulos, and A.G. Dimakis, "Optimal locally repairable codes and connections to matroid theory," *IEEE Trans. Inf. Theory,* vol. 62, no. 12, pp. 6661-6671, Dec. 2016.

[24] I. Tamo and A. Barg, "A family of optimal locally recoverable codes," *IEEE Trans. Inf. Theory,* vol. 60, no. 8, pp. 4661-4676, Aug. 2014.

[25] A. Wang and Z. Zhang, "Repair locality with multiple erasure tolerance," *IEEE Trans. Inf. Theory,* vol. 60, no. 11, pp. 6979-6987, Nov. 2014.

[26] A. Wang and Z. Zhang, "An integer programming-based bound for locally repairable codes," *IEEE Trans. Inf. Theory,* vol. 61, no. 10, pp. 5280-5294, Oct. 2015.

[27] T. Westerbäck, R. Freij-Hollanti, T. Ernvall, and C. Hollanti, "On the Combinatorics of locally repairable codes via matroid theory," *IEEE Trans. Inf. Theory,* vol. 62, no. 10, pp. 5296-5315, Oct. 2016.