How long can optimal locally repairable codes be?

Guruswami, Venkatesan; Xing, Chaoping; Yuan, Chen

2018

Guruswami, V., Xing, C., & Yuan, C. (2018). How long can optimal locally repairable codes be?. Leibniz International Proceedings in Informatics, 116, 41-. doi:10.4230/LIPIcs.APPROX-RANDOM.2018.41

https://hdl.handle.net/10356/89385

https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2018.41

How Long Can Optimal Locally Repairable Codes Be?

Venkatesan Guruswami[1]
Computer Science Department, Carnegie Mellon University, Pittsburgh, USA.
venkatg@cs.cmu.edu
https://orcid.org/0000-0001-7926-3396

Chaoping Xing
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore.
xingcp@ntu.edu.sg
https://orcid.org/0000-0002-1257-1033

Chen Yuan[2]
Centrum Wiskunde & Informatica, Amsterdam, Netherlands.
Chen.Yuan@cwi.nl
https://orcid.org/0000-0002-3730-8397

Abstract

A locally repairable code (LRC) with locality r allows for the recovery of any erased codeword symbol using only r other codeword symbols. A Singleton-type bound dictates the best possible trade-off between the dimension and distance of LRCs — an LRC attaining this trade-off is deemed *optimal*. Such optimal LRCs have been constructed over alphabets growing linearly in the block length. Unlike the classical Singleton bound, however, it was not known if such a linear growth in the alphabet size is necessary, or for that matter even if the alphabet needs to grow at all with the block length. Indeed, for small code distances $3, 4$, arbitrarily long optimal LRCs were known over fixed alphabets.

Here, we prove that for distances $d \geqslant 5$, the code length n of an optimal LRC over an alphabet of size q must be at most roughly $O(dq^3)$. For the case $d = 5$, our upper bound is $O(q^2)$. We complement these bounds by showing the existence of optimal LRCs of length $\Omega_{d,r}(q^{1+1/\lfloor(d-3)/2\rfloor})$ when $d \leqslant r + 2$. Our bounds match when $d = 5$, pinning down $n = \Theta(q^2)$ as the asymptotically largest length of an optimal LRC for this case.

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

Keywords and phrases Locally Repairable Code, Singleton Bound

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2018.41

Related Version A full version of the paper is available at `https://arxiv.org/abs/1807.01064`.

[1] This work was done when the author was visiting the School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, and the Center of Mathematical Sciences and Applications, Harvard University. Research supported in part by NSF CCF-1563742.

[2] Most of this work was done when the author was with the School of Physical and Mathematical Science, Nanyang Technological University, Singapore. Research supported in part by ERC H2020 grant No.74079 (ALGSTRONGCRYPTO).

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018).
Editors: Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer; Article No. 41; pp. 41:1–41:11
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Modern distributed storage systems have been transitioning to erasure coding based schemes with good storage efficiency in order to cope with the explosion in the amount of data stored online. Locally Repairable Codes (LRCs) have emerged as the codes of choice for many such scenarios and have been implemented in a number of large scale systems e.g., Microsoft Azure [10] and Hadoop [17].

A block code is called a locally repairable code (LRC) with locality r if every symbol in the encoding is a function of r other symbols. This enables recovery of any single erased symbol in a local fashion by downloading at most r other symbols. On the other hand, one would like the code to have a good minimum distance to enable recovery of many erasures in the worst-case. LRCs have been the subject of extensive study in recent years [8, 6, 16, 18, 11, 13, 5, 15, 19, 20, 3]. LRCs offer a good balance between very efficient erasure recovery in the typical case in distributed storage systems where a single node fails (or becomes temporarily unavailable due to maintenace or other causes), and still allowing recovery of the data from a larger number of erasures and thus safeguarding the data in more worst-case scenarios.

A Singleton-type bound for locally repairable codes relating its length n, dimension k, minimum distance d and locality r was first shown in the highly influential work [6]. It states that a linear locally repairable code C must obey[3]

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

Note that any linear code of dimension k has locality at most k, so in the case when $r = k$ the above bound specializes to the classical Singleton bound $d \leqslant n - k + 1$, and in general it quantifies how much one must back off from this bound to accommodate locality.

A linear LRC that meets the bound (1) with equality is said to be an *optimal* LRC. This work concerns the trade-off between alphabet size and code length for linear codes that are optimal LRCs. Initially, the existence of such optimal LRCs and constructions were only known over fields that were exponentially large in the block length [9, 18].[4]

In a celebrated paper, Tamo and Barg [19] constructed clever subcodes of Reed-Solomon codes that yield a class of optimal locally repairable codes inheriting the field size $q \approx n$ of Reed-Solomon codes. This shows that one can have optimal LRCs with a field size similar to that of Maximum Distance Separable (MDS) codes which attain the classical Singleton bound $d = n - k + 1$.

One is thus tempted to make an analogy between optimal LRCs and MDS codes. The famous MDS conjecture says that there are no non-trivial (meaning, distance $d > 2$) MDS codes with length exceeding $q + 1$ where q is its alphabet size, except in two corner cases (q even and $k = 3$, or $k = q - 1$) where the length is at most $q + 2$. This conjecture was famously resolved in the case when q is prime by Ball [1].

For optimal LRCs, it was shown that an analogous strong conjecture does not hold [13] for almost every distance d — using elliptic curves, they gave LRCs length $q + 2\sqrt{q}$ (an earlier construction using rational function fields achieved length $q + 1$ [11]). A construction

[3] The bound in [6] was shown even for a weaker requirement of locality only for the information symbols, but we focus on the more general all-symbol locality.

[4] If locality is desired only for the information symbols, then it is easy to construct optimal LRCs over linear-sized fields using any MDS code via the “Pyramid" construction [9]. As we said, our focus is on LRCs with all-symbol locality which is more challenging to ensure.

of length $n \approx \frac{r+1}{r}q$ was given for small distances in [21]. Note that all these constructions have length that is at most $O(q)$.

The MDS conjecture makes a very precise statement about the maximum possible length of MDS codes. An asymptotic upper bound of $n = O(q)$ (in fact even $n \leqslant 2q$) is much easier to establish for MDS codes. Given this apparent parallel and the above-mentioned constructions which don't achieve code lengths exceeding $O(q)$, one might wonder if the Tamo-Barg result is asymptotically optimal, in the sense that optimal LRCs must have length at most $O(q)$. Rather surprisingly, it was not even known if n must be bounded as a function of q at all — that is, it was conceivable that one could have arbitrarily long optimal LRCs over an alphabet of fixed size! Indeed, Barg et al. [2] gave optimal LRCs using algebraic surfaces of length $n \approx q^2$ when the distance $d = 3$ and $r \leqslant 4$. This then inspired the discovery of optimal LRCs with *unbounded length* for $d = 3, 4$ via cyclic codes [14]. In Appendix A.1, we include a simple construction of arbitrarily optimal LRCs for $d = 3, 4$ over any fixed field size that satisfies $q \geqslant r + 1$.

Our Results. Given this state of knowledge, the natural question that arises is whether there is any upper bound at all on the length of optimal locally repairable codes (as a function of its alphabet size). In this paper, we answer this question affirmatively. In fact, we show that as soon as the distance $d \geqslant 5$, one cannot have unbounded length optimal LRCs (unlike the cases of $d = 3, 4$). Below is a statement of our upper bound on the code length of optimal LRCs. To the best of our knowledge, this is the first upper bound on the length of optimal LRCs.[5]

▶ **Theorem 1** (Upper bound on code length of LRCs). *Let $d \geqslant 5$, and let C be an optimal LRC with locality r (that meets the bound (1) with equality) of length $n \geqslant \Omega(dr^2)$ over an alphabet of size q. Then $n \leqslant O(dq^3)$ when d is not divisible by 4, and $n \leqslant O(dq^{3+4/(d-4)})$ when $4|d$.*

Our actual upper bound is a bit better when $d \equiv 1 \pmod 4$ and in particular yields $n \leqslant O(q^2)$ when $d = 5$. The technical condition that n is at least $\Omega(dr^2)$ arises in ensuring that the code consists of $n/(r+1)$ *disjoint* recovery groups of size $(r + 1)$ each, that together ensure recoverability with locality r for every codeword symbols.

In our second result, we complement the above result on the limitation of LRCs with a construction of super-linear (in q) length for $d \leqslant r + 2$.

▶ **Theorem 2** (Construction of long LRCs). *For every r, d with $d \leqslant r + 2$, there exist optimal LRCs of length $n \geqslant \Omega_{d,r}(q^{1+1/\lfloor (d-3)/2 \rfloor})$.*[6]

Again, to the best of our knowledge, this is the first code achieving super linear length in q for $d \geqslant 5$. The previous best construction due to [21] achieved a length of $\frac{r+1}{r}q$ for $d \leqslant r + 1$

Organization of the paper. The paper is organized as follows. In Section 2, we provide some preliminaries on locally repairable codes. In Section 3, we prove an upper bound on

[5] Using the bound of Theorem 1 of [4], one can deduce an upper bound of $O(qr)$ on the *distance* of optimal LRCs.

[6] When $d = r + 2$, it turns out that one cannot achieve bound (1) with equality; so we get codes with $d = n - k - \lceil k/r \rceil + 1$ which is the optimal trade-off in this case. For $d \leqslant r + 1$ we attain (1) with equality.

the length of optimal LRCs. In Section 4, we present a construction of optimal LRC with super linear length in its alphabet size. Due to space restrictions, some of the proofs are omitted and can be found in the full version.

2 Preliminaries

$[n]$ stands for $\{1,\dots,n\}$. The floor function and ceiling function of x are denoted by $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively. An $[n,k,d]_q$ code is a linear code over the field of size q that has length n, dimension k, and distance d. We now define the local recoverability property of a code formally. We give this definition in general without assuming linearity, though we restrict our focus to linear codes in this paper.

▶ **Definition 3.** Let C be a q-ary block code of length n. For each $\alpha \in \mathbb{F}_q$ and $i \in \{1,2,\cdots,n\}$, define $C(i,\alpha) := \{\mathbf{c} = (c_1,\dots,c_n) \in C \ : \ c_i = \alpha\}$. For a subset $I \subseteq \{1,2,\cdots,n\} \setminus \{i\}$, we denote by $C_I(i,\alpha)$ the projection of $C(i,\alpha)$ on I. For $i \in \{1,2,\cdots,n\}$, a subset R of $\{1,2,\dots,n\}$ that contains i is a called a *recovery set* for i if $C_{I_i}(i,\alpha)$ and $C_{I_i}(i,\beta)$ are disjoint for any $\alpha \neq \beta$, where $I_i = R \setminus \{i\}$. Furthermore, C is called a locally recoverable code with locality r if, for every $i \in \{1,2,\cdots,n\}$, there exists a recovery set R_i for i of size $r+1$.

▶ Remark 4. The above definition of recovery sets is slightly different from that of recovery sets given in literature where i is excluded in the recovery set I_i. The reason why we include i in the recover set R_i of i is for convenience of proofs in this paper.

For linear codes, which are the focus of this paper, the following lemma establishes a connection between the locality and the dual code $C^{\perp}$. The proof is folklore.

▶ **Lemma 5.** *A subset R of $\{1,2,\dots,n\}$ is a recovery set at i of a linear code C over $\mathbb{F}_q$ if and only if there exists a codeword in $C^{\perp}$ whose support contains i and is a subset of R.*

For a q-ary $[n,k,d]$-linear LRC with locality r, the Singleton-type bound says

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

Like the classical Singleton bound, the Singleton-type bound (2) does not take into account the cardinality of the code alphabet q. Augmenting this result, a recent work [4] established a bound on the distance of locally repairable codes that depends on q, sometimes yielding better results. However, in this paper, we specifically refer as optimal LRC a linear code achieving the bound (2). We now rewrite this bound in a form that will be more convenient to us.

▶ **Lemma 6.** *Let n,k,d,r be positive integers with $(r+1)|n$. If the Singleton-type bound (2) is achieved, then*

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

▶ Remark 7. It turns out that the other direction of Lemma 6 is also true if $d - 2 \not\equiv r \pmod{r+1}$.

3 An Upper Bound on Code Lengths

In this section, we investigate the upper bound on the code lengths of optimal LRCs over a finite field $\mathbb{F}_q$. For simplicity, we assume that n is divisible by $r+1$ throughout this section. However, in Remark 9 and 11, we extend our results to cover the cases when n is not divisible by $r+1$.

3.1 Justifying the assumption of disjoint recovery sets

We first argue that a r-local LRC with block length n divisible by $r+1$ can be assumed, under modest conditions on the parameters, to contain $n/(r+1)$ disjoint recovery sets that each allow for recovery of $(r+1)$ codeword symbols. This structure will then be helpful to us in upper bounding the length of LRCs.

We remark that the structure theorem in [6] showed that the information symbols can be arranged into k/r disjoint groups each with a local parity check, under the assumption that $r|k$. However, we seek all-symbol's locality, and their argument does not directly apply.

▶ **Lemma 8.** *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*

$$\frac{n}{r+1} \geqslant \left(d-2-\left\lfloor\frac{d-2}{r+1}\right\rfloor\right)(3r+2)+\left\lfloor\frac{d-2}{r+1}\right\rfloor+1. \tag{4}$$

▶ Remark 9. A similar result holds when n is not divisible by $r+1$. In this case, one can guarantee $\lceil\frac{n}{r+1}\rceil$ recovery sets that cover all the n codeword positions.

3.2 Proving the upper bound

In this subsection, we prove Theorem 1 (restated more formally below) that gives an upper bound on the length n of a LRC in terms of its alphabet size q. The parity check view of an LRC will be instrumental in our argument, in a manner similar to the bound obtained for maximally recoverable (MR) LRCs in [7]. We will make use of Lemma 8 and the classical Hamming upper bound on the size of codes as a function of minimum distance to derive our result.

▶ **Theorem 10.** *Let C be an optimal $[n,k,d]_q$-linear locally repairable codes of locality r with $(r+1)|n$ and parameters satisfying the inequality* (4) *given in Lemma 8. If $d \geqslant 5$ and $d \equiv a \pmod 4$ for some $1 \leqslant a \leqslant 4$, then*

$$n = \begin{cases} O(dq^{\frac{4(d-2)}{d-a}-1}) & \text{if } a=1,2, \\ O(dq^{\frac{4(d-3)}{d-a}-1}) & \text{if } a=3,4. \end{cases} \tag{5}$$

In particular, we have $n = O\left(dq^{3+\frac{4}{d-4}}\right)$. Furthermore, we have $n = O(q^2)$, $O(q^3)$, $O(q^3)$, $O(q^4)$, $O(q^{2.5})$ and $O(q^3)$ for $d = 5,6,7,8,9,$ and 10, respectively.

Proof. Again we let $n-k = \frac{n}{r+1}+h$ with $h = d-2-\lfloor\frac{d-2}{r+1}\rfloor \leqslant d-2$. By Theorem 8, we know that there exist $\ell := \frac{n}{r+1}$ codewords $\mathbf{c}_1,\dots,\mathbf{c}_\ell$ of $C^\perp$ such that the supports $\mathrm{Supp}(\mathbf{c}_1),\dots,\mathrm{Supp}(\mathbf{c}_\ell)$, each of size $r+1$, are pairwise disjoint. Put $R_i = \mathrm{Supp}(\mathbf{c}_i)$. By considering an equivalent code, we may assume that $R_i = \{(i-1)(r+1)+1,\dots,i(r+1)\}$ for $i = 1,2,\dots,\ell$ and the projection of $\mathbf{c}_i$ at R_i are equal to all-one vector $\mathbf{1}$ of length $r+1$.

The parity-check matrix H has the following form

$$H = \left(\begin{array}{c|c|ccc|c} \mathbf{1} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{1} \\ \hdashline \multicolumn{6}{c}{A} \end{array}\right), \tag{6}$$

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 $(i(r+1)+j)$-th column of H, i.e.,

$$\mathbf{h}_{i,j} = (\underbrace{0,\dots,0}_{i-1}, 1, \underbrace{0,\dots,0}_{\ell-i}, \mathbf{v}_{i,j})^T \tag{7}$$

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

Define

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

for $i \in [\ell]$ and $j \in [r]$. We claim that any $\lfloor \frac{d-1}{2} \rfloor$ of $\mathbf{h}'_{1,1}, \dots, \mathbf{h}'_{\ell,r}$ are linearly independent. Indeed, for any $t := \lfloor \frac{d-1}{2} \rfloor$ vectors $\mathbf{h}'_{i_1,j_1}, \dots, \mathbf{h}'_{i_t,j_t}$ and scalars $\lambda_{i_1,j_1}, \dots, \lambda_{i_t,j_t} \in \mathbb{F}_q$ satisfying $\sum_{k=1}^t \lambda_{i_k,j_k} \mathbf{h}'_{i_k,j_k} = \mathbf{0}$, i.e., $\sum_{k=1}^t \lambda_{i_k,j_k}(\mathbf{h}_{i_k,j_k} - \mathbf{h}_{i_k,r+1}) = \mathbf{0}$, we have $\sum_{k=1}^t \lambda_{i_k,j_k} \mathbf{h}_{i_k,j_k} - \sum_{k=1}^t \lambda_{i_k,j_k} \mathbf{h}_{i_k,r+1} = \mathbf{0}$

Note that $\mathbf{h}_{i_1,j_1}, \dots, \mathbf{h}_{i_k,j_k}$ together with $\mathbf{h}_{i_1,r+1}, \dots, \mathbf{h}_{i_k,r+1}$ are at most $2t \leqslant d-1$ distinct columns of H. It follows that they are linearly independent and thus the coefficient $\lambda_{i_1,j_1}, \dots, \lambda_{i_t,j_t}$ must be all zero.

Moreover, we note that the first ℓ components of $\mathbf{h}'_{i,j}$ are all zero for $(i,j) \in [\ell] \times [r]$. We shorten the vector $\mathbf{h}'_{i,j}$ by puncturing its first ℓ coordinates. Denote by $\widetilde{\mathbf{h}}_{i,j}$ the shortened vectors. It is clear that any $\lfloor \frac{d-1}{2} \rfloor$ of $\widetilde{\mathbf{h}}_{1,1}, \dots, \widetilde{\mathbf{h}}_{\ell,r}$ are still linearly independent. Let H_2 be the matrix whose columns consists of $\widetilde{\mathbf{h}}_{i,j}$ for $i = 1, \dots, \ell$ and $j = 1, \dots, r$ and let C_2 be a linear code whose parity-check matrix is H_2. Then C_2 is a linear code with length $N := n - \ell = \frac{rn}{r+1}$, dimension at least $N - h$ and distance at least $\lfloor \frac{d-1}{2} \rfloor + 1$. We now apply the Hamming bound to $C_2 = \left[n - \ell, \geqslant n - \ell - h, \geqslant \lfloor \frac{d-1}{2} \rfloor + 1\right]$-linear code.

Let $d = 4d_1 + a$ for some $d_1 \geqslant 1$ and $1 \leqslant a \leqslant 4$.

Case 1. $a = 1$ or 2. In this case, we have $\lfloor \frac{d-1}{2} \rfloor + 1 = 2d_1 + 1$. Applying the Hamming bound to C_2 gives

$$q^{N-h} \leqslant \frac{q^N}{\sum_{i=1}^{d_1} \binom{N}{i}(q-1)^i} \leqslant \frac{q^N}{\binom{N}{d_1}(q-1)^{d_1}} \leqslant \frac{q^N}{(\frac{N}{d_1})^{d_1}(q-1)^{d_1}},$$

i.e., $\frac{rn}{r+1} = N \leqslant \frac{d_1}{q-1} \times q^{\frac{h}{d_1}} = \frac{d-a}{4(q-1)} \times q^{\frac{4h}{d-a}} \leqslant \frac{d-a}{4(q-1)} \times q^{\frac{4(d-2)}{d-a}}$. The last inequality follows from the fact that $h \leqslant d-2$.

Case 2. $a = 3$ or 4. In this case, we have $\lfloor \frac{d-1}{2} \rfloor + 1 = 2d_1 + 2$. Deleting the first coordinate of C_2 gives a q-ary $[N-1, N-h, \geqslant 2d_1 + 1]$-linear code. Applying the Hamming bound to $[N-1, N-h, \geqslant 2d_1 + 1]$ gives

$$q^{N-h} \leqslant \frac{q^{N-1}}{\sum_{i=1}^{d_1} \binom{N-1}{i}(q-1)^i} \leqslant \frac{q^{N-1}}{\binom{N-1}{d_1}(q-1)^{d_1}} \leqslant \frac{q^{N-1}}{(\frac{N-1}{d_1})^{d_1}(q-1)^{d_1}},$$

i.e., $\frac{rn}{r+1} - 1 = N - 1 \leqslant \frac{d_1}{q-1} \times q^{\frac{h-1}{d_1}} = \frac{d-a}{4(q-1)} \times q^{\frac{4(h-1)}{d-a}} \leqslant \frac{d-a}{4(q-1)} \times q^{\frac{4(d-3)}{d-a}}$. In conclusion, we have

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

The desired result follows. ◀

▶ Remark 11. Let us extend this result to the case n is not divisible by $r+1$. From Remark 9, we obtain $\lceil \frac{n}{r+1} \rceil$ recovery sets $R_1, \dots, R_{\lceil \frac{n}{r+1} \rceil}$ covering all of the n indices. There are at most $(r+1)\lceil \frac{n}{r+1} \rceil - n \leqslant r$ indices that belong to more than 1 of these $\lceil \frac{n}{r+1} \rceil$ recovery sets. We first build the parity-check matrix H whose first $\lceil \frac{n}{r+1} \rceil$ rows are $\mathbf{c}_1, \dots, \mathbf{c}_{\lceil \frac{n}{r+1} \rceil}$ where $\mathbf{c}_i$ corresponds to recovery set R_i. Then, we remove the columns from H whose indices belong to multiple recovery sets. After removing at most r columns, we apply the same argument to the resulting matrix. It is thus clear that the same result also holds for the case n is not divisible by $r+1$, with a small adjustment of r in the final upper bound on the code length.

▶ Remark 12. From our proof of Theorem 10, one might see why our argument is not applicable to the optimal LRC with distance less than 5. In our argument, the optimal LRC of distance d is reduced to a code of distance at least $\lfloor \frac{d+1}{2} \rfloor$ without locality. If $d \leqslant 4$, this reduced code might be the Hamming code. As we know, the length of Hamming code is independent of the alphabet size. On the other hand, there indeed exists unbounded length of optimal LRC of distance $d \leqslant 4$. Therefore, our argument reveals the inherent differences of optimal LRCS with distance less than 5 and above.

4 Construction of LRCs of super-linear length

To the best of our knowledge, all known constructions of optimal LRCs have block length $n \leqslant O(q)$ unless $d \leqslant 4$. Our upper bound in the preceding section implies that n must be upper bounded by (roughly) q^3. A natural question arises whether there exists optimal LRC with super linear length in q, e.g, $n = \Omega(q^{1+\varepsilon})$ and some constant $d > 4$. In this section we answer this question affirmatively, showing such codes for all $d \leqslant r+2$.

When $d = r+2$ and $r+1|n$, the Singleton-type bound (1) can't be met [6, Corollary 10]. In this case, by an optimal LRC we mean a code attaining the trade-off $d = n - k - \lceil \frac{k}{r} \rceil + 1$. When n is not divisible by $r+1$, by shortening the code, it is still possible to obtain the optimal LRCs.

▶ **Theorem 13.** *Assume $d \leqslant r+2$ and $(r+1)|n$. There exist optimal LRCs of length $n = \Omega_{d,r}(q^{1+\frac{1}{\lfloor (d-3)/2 \rfloor}})$. In particular, one obtains the best possible length $n = O(q^2)$ for optimal LRC of minimum distance 5 if $r \geqslant 3$ and $(r+1)|n$.*

Proof. Let $n = \eta q^{1+1/\lfloor (d-3)/2 \rfloor}$ with some constant η that only depends on d and r, i.e., $\eta = \Omega_{d,r}(1)$. We will determine η later. It suffices to construct a matrix H and show that the code C derived from this parity-check matrix is an optimal LRC. Label and order the n coordinates with $(i,j) \in [\frac{n}{r+1}] \times [r+1]$, i.e., (i_1, j_1) precedes (i_2, j_2) if $i_1 < i_2$ or $i_1 = i_2$ and $j_1 < j_2$. Let $H = (\mathbf{h}_{i,j})_{(i,j) \in [\frac{n}{r+1}] \times [r+1]}$ where $\mathbf{h}_{i,j} \in \mathbb{F}_q^{n-k}$. That means H consists of the columns $\mathbf{h}_{i,j}$ for $(i,j) \in [\frac{n}{r+1}] \times [r+1]$. We start from $\mathbf{h}_{1,1}$ and determine the value of $\mathbf{h}_{i,j}$ column by column in the above order. In each step, we make sure that the new column $\mathbf{h}_{i,j}$ together with any $d-2$ columns preceding the (i,j)-th column are linearly independent. Meanwhile, the matrix H holds the same form[7] as the matrix in (6). If we can achieve both of the conditions, we are done. Define $\frac{n}{r+1}$ blocks $B_1, \dots, B_{\frac{n}{r+1}}$ such that $B_i = \{\mathbf{h}_{i,1}, \dots, \mathbf{h}_{i,r+1}\}$. That means we partition the n columns into $\frac{n}{r+1}$ disjoint blocks. Algorithm 1 gives the iterative method to compute the columns $\mathbf{h}_{i,j}$'s.

[7] The same form is referred to that their distributions of non-zero entry in upper half matrix (matrix lying above A) are the same, i.e., entries of value 1 and 0 in this upper half matrix represents the nonzero and zero entries, respectively.

Algorithm 1

```
- For $i = 1, \dots, \frac{n}{r+1}$, and $j = 1, \dots, r+1$, do the following operation.
    - Find $\mathbf{v} \in \mathbb{F}_q^{n-k}$ of form (7)[8] such that $\mathbf{v}$ is linearly independent of any subset of at
      most $(d-2)$ columns $\mathbf{h}_{i,j}$ chosen before this step.
    - Let $\mathbf{v}$ be the $(i,j)$-th column of $H$, i.e., $\mathbf{h}_{i,j} = \mathbf{v}$.
```

We justify Algorithm 1 by showing that there always exists such $\mathbf{h}_{i,j}$ for any $(i,j) \in [\frac{n}{r+1}] \times [r+1]$. Assume that we arrive at the (a,b)-th column. If $b = 1$, the construction is trivial. Let $\mathbf{h}_{a,b}$ be a column vector such that the first $\frac{n}{r+1}$ components except i-th component are zero. Obviously, it matches the form of Equation 7. The linearly independence is also trivial since the i-th component of all the columns $\mathbf{h}_{i,j}$ for $i < a$ is 0. Otherwise, to simplify our discussion, we assume that the first $d-2$ columns are already found. Since any $d-2$ columns prior to the (a,b)-th column are already linearly independent by our algorithm, it suffices to show that $\mathbf{h}_{a,b}$ is linearly independent from these $d-2$ columns. To achieve this, we need to check all possible combinations of these $d-2$ columns. Assume that these $d-2$ columns are chosen exactly from t blocks. Obviously, block B_a must be selected. Otherwise, the same reason for $b = 1$ implies that $\mathbf{h}_{a,b}$ is linearly independent of these $d-2$ columns. Without loss of generality, we assume that these t blocks are $B_1, \dots, B_{t-1}$ and B_a and there are i_j columns picked from block B_j. Then, the submatrix H_1 consisting of these $d-2$ columns has the following form:

$$H_1 = \left(\begin{array}{c|c|c|c|c}
\mathbf{x}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{0} & \mathbf{0} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{t-1} & \mathbf{0} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_a \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\hdashline
\multicolumn{5}{c}{A_1}
\end{array}\right)$$

where $\mathbf{x}_i \in \mathbb{F}_q^{i_j}$ and A_1 is a $(d-2) \times (d-2)$ matrix. If any B_j, $j = 1, 2, \dots, t-1$, contains only one column, then that column is linearly independent of the rest of the $d-3$ columns and $\mathbf{h}_{a,b}$, and therefore can be removed from consideration. Thus we may assume that there are at least two columns chosen in each block except block B_a. Thus, t is at most $\lfloor \frac{d-1}{2} \rfloor$. Recall that our goal is to ensure that $\mathbf{h}_{a,b}$ is linearly independent of these at most $d-2$ columns in total chosen from the blocks $B_1, \dots, B_{t-1}$. Given the t blocks and the $d-2$ columns chosen from them, we count the number of bad $\mathbf{h}_{a,b}$ which are linear combinations of these $d-2$ columns. If the number of such linear combinations is smaller than the size of the whole space of possible choices of $\mathbf{h}_{a,b}$, we are done. To achieve this, we need to determine the maximal subspace V spanned by these $d-2$ columns such that all the vectors in V has the same form as the vector $\mathbf{h}_{a,b}$, i.e., the first $\frac{n}{r+1}$ components except a-th component are zero.

For block B_j with $j \neq a$, by the expression of matrix H_1, the i_j columns of B_j created a $i_j - 1$-dimensional subspace where the first $\frac{n}{r+1}$ components of all the vectors are 0. That

[8] Only the i-th component out of the first $\frac{n}{r+1}$ components is nonzero.

means, block B_j for $j \neq a$ contributes $i_j - 1$ linearly independent vectors to the maximal subspace V. For block B_a, it contributes at most i_a linearly independent vectors to the maximal subspace V. It follows that the dimension of V is at most $\sum_{i=1}^{t-1}(i_j-1)+i_a = d-1-t$. This implies that there are at most q^{d-1-t} $\mathbf{h}_{a,b}$s lying in the space spanned by these $d-2$ columns.

It remains to count the number of distinct $d-2$ column sets. Note that B_a is always selected. Thus, we only have at most $\binom{a-1}{t-1} \leqslant \binom{\frac{n}{r+1}}{t-1} \leqslant (\frac{n}{r+1})^{t-1}$ combinations of these t blocks. After fixing these t blocks, there are at most $(t(r+1))^{d-2}$ ways to pick $d-2$ columns from these t blocks due to the fact that these t blocks contain only $t(r+1)$ columns. In total, there are at most $(t(r+1))^{d-2}(\frac{n}{r+1})^{t-1}$ ways to pick $d-2$ columns that precede the (i,j)-th column. Each combination contributes to at most q^{d-1-t} bad $\mathbf{h}_{a,b}$. Thus, the number of bad $\mathbf{h}_{a,b}$ are upper bounded by

$$\begin{aligned}\sum_{t=1}^{\lfloor\frac{d-1}{2}\rfloor}\big(t(r+1)\big)^{d-2}\left(\frac{n}{r+1}\right)^{t-1}q^{d-1-t} &\leqslant \left(\frac{q(d-1)(r+1)}{2}\right)^{d-2}\sum_{t=1}^{\lfloor\frac{d-1}{2}\rfloor}\left(\frac{n}{q(r+1)}\right)^{t-1}\\ &\leqslant \left(\frac{q(d-1)(r+1)}{2}\right)^{d-2}\left(\frac{d-1}{2}\right)\left(\frac{n}{q(r+1)}\right)^{\lfloor\frac{d-3}{2}\rfloor}.\end{aligned}$$

The first inequality is due to $t \leqslant \frac{d-1}{2}$ and the last inequality is due to $n > q(r+1)$. Plug $n = \eta q^{1+\frac{1}{\lfloor(d-3)/2\rfloor}}$ into the formula. This number is upper bounded by $q^{d-1}(d-1)^{d-1}(r+1)^{(d-2)/2}\eta^{\lfloor(d-3)/2\rfloor}$; by picking η small enough as a function of d, r we can ensure this quantity is at most $q^{d-1}/2$.

On the other hand, according to Algorithm 1, the a-th component of $\mathbf{h}_{a,b}$ should be nonzero. Moreover, the first $\frac{n}{r+1}$ components except a-th component are all zero. That means, the whole space of $\mathbf{h}_{a,b}$ is of size $q^{d-1} - q^{d-2} > \frac{1}{2}q^{d-1}$. Thus, there always exists $\mathbf{h}_{a,b}$ satisfying our algorithm's requirement.

We are almost done. Let C be the code whose parity-check matrix is H. It is clear that C has locality r. Since any $d-1$ columns of H are linearly independent, C has minimum distance $d(C)$ at least d. Because H has $\frac{n}{r+1} + d - 2$ rows, the dimension of C is $k(C) \geqslant n - \frac{n}{r+1} - (d-2) = \frac{rn}{r+1} - (d-2)$. This implies

$$\frac{k(C)}{r} \geqslant \frac{n}{r+1} - \frac{d-2}{r}$$

We divide it into two case.

- If $d-2 < r$, the condition $r+1|n$ implies $\left\lceil\frac{k(C)}{r}\right\rceil \geqslant \frac{n}{r+1}$ and thus $k(C)+\left\lceil\frac{k(C)}{r}\right\rceil \geqslant n-d+2$. It follows that

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

  Thus, C is an optimal LRC. We are done.
- If $d-2 = r$, the condition $r+1|n$ implies $\left\lceil\frac{k(C)}{r}\right\rceil \geqslant \frac{n}{r+1} - 1$ and thus $k(C) + \left\lceil\frac{k(C)}{r}\right\rceil \geqslant n - d + 1$. It follows that

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

  C is still an optimal LRC because there does not exist LRC reaching the Singleton-type bound. ◀

Next, we extend this theorem to the case n is not divisible by $r+1$ and $d \leqslant r+2$.

▶ **Corollary 14.** *Assume* $n \equiv a \pmod{r+1}$ *and* $a > d-1$. *There exists optimal LRC of length* $n = \Omega_{d,r}(q^{1+\frac{1}{\lfloor (d-3)/2 \rfloor}})$. *In particular, one obtains the best possible length* $n = O(q^2)$ *for optimal LRC of minimum distance* 5 *if* $n \pmod{r+1} > 4$.

It was shown in [12, Theorem III.3] that under the assumption that C has $\lceil \frac{n}{r+1} \rceil$ disjoint recovery sets, a linear code C with length $n = a \bmod r+1$, $a \neq 0, 1$ and dimension either $k \bmod r \geqslant a$ or $r|k$ must obey that $d \leqslant n - k - \lceil \frac{k}{r} \rceil + 1$.

With the help of Theorem III.3 in [12], we can extend the result in Corollary 14 to cover the case $a \leqslant d-1$ and $a \neq 1$.

▶ **Corollary 15.** *Assume* $n \equiv a \pmod{r+1}$ *and* $a \neq 1$. *There exists optimal LRC of length* $n = \Omega_{d,r}(q^{1+\frac{1}{\lfloor (d-3)/2 \rfloor}})$.

References

1 S. Ball. On large subsets of a finite vector space in which every subset of basis size is a basis. *J. Eur*, 14:733–748, October 2012.

2 A. Barg, K. Haymaker, E. Howe, G. Matthews, and A. Várilly-Alvarado. Locally recoverable codes from algebraic curves and surfaces. In E. W. Howe, K. E. Lauter, and J. L. Walker, editors, *Algebraic Geometry for Coding Theory and Cryptography*, pages 95–126. s, Springer, 2017.

3 A. Barg, I. Tamo, and S. Vlăduţ. Locally recoverable codes on algebraic curves. *IEEE Trans. Inform. Theory*, 63:4928–4939, 2017.

4 V. Cadambe and A. Mazumda. Bounds on the size of locally recoverable codes. *IEEE Trans. Inform. Theory*, 61:5787–5794, 2015.

5 M. Forbes and S. Yekhanin. On the locality of codeword symbols in non-linear codes. *Discrete Mathematics*, 324(6):78–84, 2014.

6 P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin. On the locality of codeword symbols. *IEEE Trans. Inform. Theory*, 58:6925–6934, 2012.

7 S. Gopi, V. Guruswami, and S. Yekhanin. On maximally recoverable local reconstruction codes. *Electronic Colloquium on Computational Complexity*, 24::183, 2017.

8 J. Han and L. A. Lastras-Montano. Reliable memories with subline accesses. In *Proc. IEEE Internat. Sympos. Inform. Theory*, pages 2531–2535, 2007.

9 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*, pages 79–86, 2007.

10 C. Huang, H. Simitci, Y. Xu, Ogus A., B. Calder, P. Gopalan, J. Li, and S. Yekhanin. Erasure coding in windows azure storage. *In USENIX Annual Technical Conference (ATC)*, pages 15–26, 2012.

11 L. Jin, L. Ma, and Xing C. Construction of optimal locally repairable codes via automorphism groups of rational function fields. URL: `https://arxiv.org/abs/1710.09638`.

12 O. Kolosov, A. Barg, I. Tamo, and G. Yadgar. Optimal lrc codes for all lengths $n \leqslant q$. URL: `https://arxiv.org/pdf/1802.00157`.

13 X. Li, L. Ma, and C. Xing. *Optimal locally repairable codes via elliptic curves.* To appear in IEEE Trans. Inf. Theory, 2017. URL: `https://arxiv.org/abs/1712.03744`.

14 Y. Luo, C. Xing, and C. Yuan. *Optimal locally repairable codes of distance* 3 *and* 4 *via cyclic codes.* To appear in IEEE Trans. Inf. Theory, 2018. `arXiv:1801.03623`.

15 D. S. Papailiopoulos and A. G. Dimakis. Locally repairable codes. *IEEE Trans. Inform. Theory*, 60:5843–5855, 2014.

16 N. Prakash, G. M. Kamath, V. Lalitha, and P. V. Kumar. Optimal linear codes with a local-error-correction property. In *Proc. 2012 IEEE Int. Symp. Inform. Theory*, pages 2776–2780, 2012.

17 M. Sathiamoorthy, M. Asteris, D. S. Papailiopoulos, A. G. Dimakis, R. Vadali, S. Chen, and D. Borthakur. XORing elephants: novel erasure codes for big data. *Proceedings of VLDB Endowment (PVLDB)*, pages 325–336, 2013.

18 N. Silberstein, A. S. Rawat, O. O. Koyluoglu, and S. Vichwanath. Optimal locally repairable codes via rank-matric codes. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 1819–1823, 2013.

19 I. Tamo and A. Barg. A family of optimal locally recoverable codes. *IEEE Trans. Inform. Theory*, 60:4661–4676, 2014.

20 I. Tamo, D. S. Papailiopoulos, and A. G. Dimakis. Optimal locally repairable codes and connections to matroid theory. *IEEE Trans. Inform. Theory*, 62:6661–6671, 2016.

21 Z. Zhang, J. Xu, and M. Liu. Constructions of optimal locally repairable codes over small fields. *SCIENTIA SINICA Mathematica*, 47(11):1607–1614, 2017.

A Appendix

A.1 Unbounded length LRCs for distances 3 and 4

▶ **Theorem 16.** *Assume that $d = 3, 4$, $d-2 \leqslant r$ and $r+1|n$, there exist optimal LRCs of arbitrarily lengths as long as $q \geqslant r+1$.*

Proof. For $d = 3, 4$, $d-2 \leqslant r$ and $r+1|n$, the Singleton-type bound implies that $n - k = \frac{n}{r+1} + d - 2$. Since $q \geqslant r+1$, we let A be a $(d-2) \times (r+1)$ Vandermonde matrix over $\mathbb{F}_q$ such that

$$A_1 = \begin{pmatrix} \mathbf{1} \\ A \end{pmatrix}$$

is a $(d-1) \times (r+1)$ Vandermonde matrix. Define $(\frac{n}{r+1} + d - 2) \times n$ matrix

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

Where $\mathbf{1}$ and $\mathbf{0}$ are all-1 and all-0 vectors in $\mathbb{F}_q^{r+1}$, respectively. We partition the columns of H into $\frac{n}{r+1}$ blocks $B_1, \dots, B_{\frac{n}{r+1}}$ such that $\mathbf{h} \in B_i$ if its i-th component is non-zero. From the expression of matrix H, it is clear that each column belongs to exactly one block and the columns in distinct blocks are linearly independent. Moreover, any $d-1$ columns in the same block are linearly independent due to the property of Vandermonde matrix A_1. Next we show that any $d-1$ columns of H are linearly independent. It suffices to verify this claim for the case $d = 4$. To see this, we pick any three columns $\mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_t$ from H. Nothing needs to prove if these three columns belong to the same block. We assume that they belong to at least two blocks. Without loss of generality, $\mathbf{h}_t$ is in a block that does not contain $\mathbf{h}_i$ and $\mathbf{h}_j$. From above observation, we see that $\mathbf{h}_t$ is linearly independent from $\mathbf{h}_i$ and $\mathbf{h}_j$. It is clear $\mathbf{h}_i$ and $\mathbf{h}_j$ are linearly independent no matter whether they belong to the same block or different blocks. Thus, any 3 columns of H are linearly independent. Let C be the linear code whose parity-check matrix is H. It is clear that C has length n, dimension $k(C) \geqslant n - \frac{n}{r+1} - (d-2) = \frac{rn}{(r+1)} - (d-2)$, distance $d(C) \geqslant d$ and locality r. The condition $d - 2 \leqslant r$ leads to $\lceil \frac{k(C)}{r} \rceil \geqslant \frac{n}{r+1}$ and thus $k(C) + \lceil \frac{k(C)}{r} \rceil \geqslant n - (d-2)$. The desired result follows since $d(C) \geqslant d \geqslant n - k(C) - \lceil \frac{k(C)}{r} \rceil + 2$. ◀