Departament d'Enginyeria Telemàtica
UNIVERSITAT POLITÈCNICA DE CATALUNYA

PhD Research Plan

LOCALLY RECOVERABLE CODES APPLICATIONS

Petar Hlad Colic

Information Security Group
Department of Network Engineering
Universitat Politècnica de Catalunya

Advisor: Marcel Fernández Muñoz

Barcelona, July 2018

CONTENTS

CHAPTER 1

CONTEXT AND MOTIVATION

In recent years the explosion in the volumes of data being stored online has resulted in distributed storage systems transitioning to erasure coding based schemes in order to ensure reliability with low storage overheads. On such a massive scale, unreachable or failed servers are no longer an exception but a regular occurrence and recovery from such events has to be done efficiently.

In recent years Locally Recoverable Codes (LRC) emerged as the codes of choice for many such scenarios and have been implemented in a number of large scale systems ([8], [11]).

Classical erasure correcting codes guarantee that data can be recovered if a bounded number of codeword coordinates is erased. However recovering data typically involves accessing all surviving coordinates. LRC codes have the property that a symbol of the codeword can be recovering accessing few other symbols of the codeword (called the *recovering set*.

Symbols can have more than one recovering set. Having more than one recovering set is beneficial in practice because it enables more users to access a given portion of data, thus enhancing data availability in the system.

Consider a linear $[n, k, d]_q$ code $\mathcal{C} \subset \mathbb{F}_q^n$. We say that the i-th coordinate of $\mathcal{C}$ has locality r, if the value at this coordinate can be recovered from accessing some other r coordinates of $\mathcal{C}$. We say that the code $\mathcal{C}$ has locality r if every symbol of the codeword $x \in \mathcal{C}$ can be recovered from a subset of r other symbols of x.

Data storage applications require codes with small redundancy, low locality for information coordinates, large distance, and low locality for parity coordinates.

Definition 1.0.1 (LRC Codes)**.** A code $\mathcal{C} \subset \mathbb{F}_q^n$ is LRC with locality r if for every $i \in [n] := \{1, 2, ..., n\}$ there exists a subset $\mathcal{R}_i \subset [n] \setminus \{i\}$, $|\mathcal{R}_i| \leq r$ and a function ϕ_i such that for every codeword $\mathbf{x} \in \mathcal{C}$ we have

$$\mathbf{x}_i = \phi_i(\{\mathbf{x}_j,\ j \in \mathcal{R}_i\}) \tag{1.1}$$

This definition can be also rephrased as follows. Given $a \in \mathbb{F}_q$ consider the sets of codewords

$$\mathcal{C}(i, a) = \{x \in \mathcal{C} : x_i = a\}, \quad i \in [n]$$

The code $\mathcal{C}$ is said to have locality r if for every $i \in [n]$ there exists a subset $\mathcal{R}_i \subset [n] \setminus \{i\},\ |\mathcal{R}_i| \leq r$ such that the restrictions of the sets $\mathcal{C}(i,a)$ to the coordinates in $\mathcal{R}_i$ for different a are disjoint:

$$\mathcal{C}_{I_i}(i,a) \cap \mathcal{C}_{I_i}(i,a') = \emptyset, \quad a \neq a' \tag{1.2}$$

The subset I_i is called a *recovering set* for the symbol x_i.

Definition 1.0.2 (t-LRC Codes)**.** A code $\mathcal{C}$ is said to have t disjoint recovering sets if for every $i \in [n]$ there are pairwise disjoint subsets $R_i^1, ..., R_i^t \subset [n] \setminus \{i\}$ such that for all $j = 1, ..., t$ and every pair of symbols $a, a' \in \mathbb{F}_q,\ a \neq a'$

$$\mathcal{C}(i,a)_{R_i^j} \cap \mathcal{C}(i,a')_{R_i^j} = \emptyset \tag{1.3}$$

Let $\mathcal{C}$ be an (n,k,r) LRC code. The rate of $\mathcal{C}$ satisfies

$$\frac{k}{n} \leq \frac{r}{r+1} \tag{1.4}$$

The minimum distance of $\mathcal{C}$ satisfies [4]

$$d \leq n - k - \left\lceil \frac{k}{r} \right\rceil + 2 \tag{1.5}$$

For (n,k,r,t) LRC codes with $t \geq 2$ disjoint recovering sets [10]:

$$d \leq n - k + 2 - \left\lceil \frac{t(k-1)+1}{t(r-1)+1} \right\rceil \tag{1.6}$$

We will refer to codes attaining the bound above as optimal LRC codes.

Let $C \in \mathbb{F}_q^{k \times n}$. The encoding of $\mathbf{x} \in \mathbb{F}_q^k$ is given by $\mathcal{C}(\mathbf{x}) = \mathbf{x}^T \cdot C$. Thus the code $\mathcal{C}$ is determined by the set of n points $C = \{\mathbf{c}_1, ..., \mathbf{c}_n\} \subset \mathbb{F}_q^k$

C must have full rank for $\mathcal{C}$ to have k information symbols.

The code $\mathcal{C}$ has distance d if and only if for every $S \subseteq C$ such that $\text{Rank}(S) \leq k-1$,

$$|S| \leq n - d \tag{1.7}$$

CHAPTER 2

STATE OF THE ART

CHAPTER 3

OPEN PROBLEMS

[5]

3.1 Constructions of binary LRC codes

[14]

3.2 Constructions of codes on algebraic curves

[3, 2, 6]

3.3 Bounds on codes with locality

[1] [12] [13]

Theorem 3.3.1 ([12])**.** *Let $\mathcal{C}$ be an (n, k, r, t) LRC code with t disjoint recovering sets of size r. Then the rate of C satisfies*

$$\frac{k}{n} \leq \frac{1}{\prod_{j=1}^{t}(1 + \frac{1}{jr})} \tag{3.1}$$

The minimum distance of C is bounded above as follows:

$$d \leq n - \sum_{i=0}^{t} \left\lfloor \frac{k-1}{r^i} \right\rfloor \tag{3.2}$$

3.4 List decoding of LRC codes

A code of length n is called (τ, ℓ)-list decodable if the Hamming sphere of radius τ centered at any vector v of length n always contains at most ℓ codewords $c \in \mathcal{C}$. It was shown by Johnson in [9] that any code of length n and distance d is (τ_J, ℓ)-list decodable where $\tau_J = n - \sqrt{n(n-d)}$ and $\ell \in poly(n)$.

It was recently shown by Holzbaur and Wachter-Zeh in [7] that

BIBLIOGRAPHY

[1] Abhishek Agarwal, Alexander Barg, Sihuang Hu, Arya Mazumdar, and Itzhak Tamo. "Combinatorial Alphabet-Dependent Bounds for Locally Recoverable Codes". In: *CoRR* abs/1702.02685 (2017). arXiv: 1702.02685. URL: http://arxiv.org/abs/1702.02685 (cit. on p. 5).

[2] Alexander Barg, Kathryn Haymaker, Everett W. Howe, Gretchen L. Matthews, and Anthony Várilly-Alvarado. "Locally recoverable codes from algebraic curves and surfaces". In: *CoRR* abs/1701.05212 (2017). arXiv: 1701.05212. URL: http://arxiv.org/abs/1701.05212 (cit. on p. 5).

[3] Alexander Barg, Itzhak Tamo, and Serge Vladuts. "Locally recoverable codes on algebraic curves". In: *CoRR* abs/1603.08876 (2016). arXiv: 1603.08876. URL: http://arxiv.org/abs/1603.08876 (cit. on p. 5).

[4] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin. "On the Locality of Codeword Symbols". In: *IEEE Transactions on Information Theory* 58.11 (2012), pp. 6925–6934. ISSN: 0018-9448. DOI: 10.1109/TIT.2012.2208937 (cit. on p. 2).

[5] Sivakanth Gopi, Venkatesan Guruswami, and Sergey Yekhanin. "On Maximally Recoverable Local Reconstruction Codes". In: *CoRR* abs/1710.10322 (2017). arXiv: 1710.10322. URL: http://arxiv.org/abs/1710.10322 (cit. on p. 5).

[6] Kathryn Haymaker, Beth Malmskog, and Gretchen Matthews. "Locally Recoverable Codes with Availability $t \geq 2$ from Fiber Products of Curves". In: abs/1612.03841 (2016). arXiv: 1612.03841. URL: http://arxiv.org/abs/1612.03841 (cit. on p. 5).

[7] Lukas Holzbaur and Antonia Wachter-Zeh. "List Decoding of Locally Repairable Codes". In: *CoRR* abs/1801.04229 (2018). arXiv: 1801.04229. URL: http://arxiv.org/abs/1801.04229 (cit. on p. 5).

[8] Cheng Huang, Huseyin Simitci, Yikang Xu, Aaron Ogus, Brad Calder, Parikshit Gopalan, Jin Li, and Sergey Yekhanin. “Erasure Coding in Windows Azure Storage”. In: *Presented as part of the 2012 USENIX Annual Technical Conference (USENIX ATC 12)*. Boston, MA: USENIX, 2012, pp. 15–26. ISBN: 978-931971-93-5. URL: https://www.usenix.org/conference/atc12/technical-sessions/presentation/huang (cit. on p. 1).

[9] S. Johnson. “A new upper bound for error-correcting codes”. In: *IRE Transactions on Information Theory* 8.3 (1962), pp. 203–207. ISSN: 0096-1000. DOI: 10.1109/TIT.1962.1057714 (cit. on p. 5).

[10] A. S. Rawat, D. S. Papailiopoulos, A. G. Dimakis, and S. Vishwanath. “Locality and Availability in Distributed Storage”. In: *IEEE Transactions on Information Theory* 62.8 (2016), pp. 4481–4493. ISSN: 0018-9448. DOI: 10.1109/TIT.2016.2524510 (cit. on p. 2).

[11] Maheswaran Sathiamoorthy, Megasthenis Asteris, Dimitris Papailiopoulos, Alexandros G. Dimakis, Ramkumar Vadali, Scott Chen, and Dhruba Borthakur. “XORing Elephants: Novel Erasure Codes for Big Data”. In: *Proc. VLDB Endow.* 6.5 (Mar. 2013), pp. 325–336. ISSN: 2150-8097. DOI: 10.14778/2535573.2488339. URL: http://dx.doi.org/10.14778/2535573.2488339 (cit. on p. 1).

[12] I. Tamo, A. Barg, and A. Frolov. “Bounds on the Parameters of Locally Recoverable Codes”. In: *IEEE Transactions on Information Theory* 62.6 (2016), pp. 3070–3083. ISSN: 0018-9448. DOI: 10.1109/TIT.2016.2518663 (cit. on p. 5).

[13] Itzhak Tamo and Alexander Barg. “Bounds on Locally Recoverable Codes with Multiple Recovering Sets”. In: *CoRR* abs/1402.0916 (2014). arXiv: 1402.0916. URL: http://arxiv.org/abs/1402.0916 (cit. on p. 5).

[14] Itzhak Tamo, Alexander Barg, Sreechakra Goparaju, and A. Robert Calderbank. “Cyclic LRC Codes, binary LRC codes, and upper bounds on the distance of cyclic codes”. In: *CoRR* abs/1603.08878 (2016). arXiv: 1603.08878. URL: http://arxiv.org/abs/1603.08878 (cit. on p. 5).