Locally Recoverable Codes

PhD Research Plan

Petar Hlad Colic

Department of Network Engineering
Universitat Politècnica de Catalunya

July 2018

Introduction

- Fast growth of data stored online.
- Distributed Storage Systems schemes based on erasure coding.
- Typical situation: unreachable or failed server.
- Recovery needs to be efficient.
- Classical Erasure Correcting codes: recovering data typically involves accessing all surviving coordinates.
- LRC Codes: any symbol can be recovered accessing few other symbols.

State of the Art

Given $a \in A$ consider the set of codewords with fixed value a at coordinate i:

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

For $I \subseteq [n]$ let $\mathcal{C}_I$ be the restriction of the code $\mathcal{C}$ to the coordinate set I:

$$\mathcal{C}_I := \{(x_i)_{i \in I} | (x_1, ..., x_n) \in \mathcal{C}\}$$

Definition 2.1.

A code $\mathcal{C}$ of length n has **locality r** *if*

$$\forall i \in [n], \quad \exists I_i \subseteq [n] \setminus \{i\}, \quad |I_i| \leq r \quad s.t.$$

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

Definition 2.2.

A code $\mathcal{C}$ of length n is said to **have** *t* **disjoint recovering sets** *if*

$\forall i \in [n]$, $\exists R_i^1, ..., R_i^t \subset [n] \setminus \{i\}$ *pairwise disjoint subsets s.t.*

$$\mathcal{C}_{R_i^j}(i, a) \cap \mathcal{C}_{R_i^j}(i, a') = \emptyset, \quad a \neq a', \ \forall j \in [t]$$

- Let $\mathcal{C}$ be a linear code of length n and dimension k.
- We say $\mathcal{C}$ is an (n, k, r)-LRC code if it has locality r and has a single recovering set for each coordinate.
- We say $\mathcal{C}$ is an (n, k, r, t)-LRC code if it has $t \geq 2$ recovering sets for each coordinate.

Bounds on LRC parameters

Let $\mathcal{C}$ be an (n, k, r) LRC code. Then:

Theorem 2.3 (Upper bound on the rate).

The rate of $\mathcal{C}$ satisfies

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

Theorem 2.4 (Extension of Singleton bound).

The minimum distance of $\mathcal{C}$ satisfies

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

If equality holds, we call $\mathcal{C}$ an **optimal LRC code**.

Bounds on LRC parameters

Let $\mathcal{C}$ be an (n, k, r, t) LRC code. Then:

Theorem 2.5 (Extension of Singleton bound for $t \geq 2$).

The minimum distance of $\mathcal{C}$ is bounded as follows

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

In particular, for $t = 1$, the bound matches the one in Theorem 2.4.

Bounds on LRC parameters

Let $\mathcal{C}$ be an (n, k, r, t) LRC code. Then:

Theorem 2.6.

The rate of $\mathcal{C}$ satisfies

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

Theorem 2.7.

The minimum distance of C is bounded above as follows

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

Construction of optimal LRC code

We want to construct an optimal (n, k, r)-LRC code.

Assume $r|k$ and $(r+1)|n$.

We need:

- $A_1, \ldots, A_{\frac{n}{r+1}} \subset \mathbb{F}_q$ disjoint subsets of size $r+1$
- $g(x) \in \mathbb{F}_q[x]$ a polynomial s.t.
  1. $deg(g) = r+1$
  2. g is constant on each set A_i: $g(\alpha) = g(\beta)$ for $\alpha, \beta \in A_i$

We will call g a good polynomial.

Construction of LRC codes

Let $A = \bigcup_{i=1}^{\frac{n}{r+1}} A_i \subseteq \mathbb{F}_q$, $|A| = n$.

Write message vectors $a \in \mathbb{F}_q^k$ as $r \times \frac{k}{r}$ matrices.

$$a = \begin{pmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,\frac{k}{r}-1} \\ a_{1,0} & a_{1,1} & \cdots & a_{1,\frac{k}{r}-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r-1,0} & a_{r-1,1} & \cdots & a_{r-1,\frac{k}{r}-1} \end{pmatrix}$$

Construction of LRC codes

Encoding polynomial

Given message vector $a \in \mathbb{F}_q^k$, define **encoding polynomial** as:

$$f_a(x) = \sum_{i=0}^{r-1} \sum_{j=0}^{\frac{k}{r}-1} a_{ij} \cdot x^i \cdot g(x)^j$$

The codeword for $a \in \mathbb{F}_q^k$ is $(f_a(\alpha))_{\alpha \in A}$

LRC code

The (n, k, r) LRC code $\mathcal{C}$ is defined as the set of n-dimensional vectors

$$\mathcal{C} = \{(f_a(\alpha), \alpha \in A) : a \in \mathbb{F}_q^k\}$$

Remark 2.8.

$$x \in A_\ell \quad \Rightarrow \quad g(x) \text{ constant}$$

$$\Rightarrow \quad \sum_{j=0}^{\frac{k}{r}-1} a_{ij} g(x)^j \text{ constant}$$

$$\Rightarrow \quad deg(f_a(x)) = deg(\sum_{i=0}^{r-1} \sum_{j=0}^{\frac{k}{r}-1} a_{ij} x^i g(x)^j) \leq r-1$$

Recovery of the erased symbol

Suppose erased symbol: $\alpha \in A_j$.
Let $(c_\beta,\ \beta \in A_j \setminus \alpha)$ denote the remaining r symbols of the recovering set.
To find the value $c_\alpha = f_a(\alpha)$, find the unique polynomial $\delta(x)$ s.t.

- $\deg(\delta(x)) \leq r$
- $\delta(\beta) = c_\beta \quad \forall \beta \in A_j \setminus \alpha$

This polynomial is:

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

Finally, set $c_\alpha = \delta(\alpha)$.

Theorem 2.9.

The linear code $\mathcal{C}$ defined has dimension k and is an optimal (n, k, r) LRC code.

Proof of dimension.

For $i \in \{0, \ldots, r-1\}$; $j \in \{0, \ldots, \frac{k}{r-1}\}$ the k polynomials $x^i g(x)^j$ are linearly independent over $\mathbb{F}$.

$\Rightarrow$ The mapping $a \mapsto f_a$ is injective.

$$\deg(f_a(x)) \leq \deg(x^{r-1} g(x)^{\frac{k}{r}-1}) = r - 1 + (r+1)(\frac{k}{r} - 1)$$

$$= k + \frac{k}{r} - 2$$

The dimension of the code is k if $n > \deg(f_a(x))$
Same as minimum distance $d(\mathcal{C}) > 0$. □

Proof of optimality.

Since the encoding is linear:

$$d(\mathcal{C}) \geq n - \max_{a \in \mathbb{F}_q^k} \deg(f_a) = n - k - \frac{k}{r} + 2$$

But we have that $d(\mathcal{C}) \leq n - k - \frac{k}{r} + 2$. Therefore, we have equality and thus it is an optimal LRC Code. □

Example: (9,4,2) LRC code

We will now construct a $(n = 9, k = 4, r = 2)$ LRC code over the field $\mathbb{F}_q$.

$$q = |\mathbb{F}_q| \geq n \quad \Rightarrow \quad q \geq 9$$

Choose $q = 13$

$$\mathcal{A} = \{A_1 = \{1, 3, 9\}, A_2 = \{2, 6, 5\}, A_3 = \{4, 12, 10\}\}$$

.

$$g(x) = x^3 = \begin{cases} 1 & \text{if } x \in A_1 \\ 8 & \text{if } x \in A_2 \\ 12 & \text{if } x \in A_3 \end{cases}$$

For $a = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \in \mathbb{F}_{13}^4$ define the encoding polynomial:

$$f_a(x) = \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \begin{pmatrix} 1 \\ x^3 \end{pmatrix} = a_{00} + a_{10}x + a_{01}x^3 + a_{11}x^4$$

E.g. $a = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \Rightarrow \quad f_a(x) = 1 + x + x^3 + x^4$

$$c = (f_a(1), f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$= (4, 8, 7, 1, 11, 2, 0, 0, 0)$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(4, 8, 7, 1, 11, 2, 0, 0, 0)$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(\cancel{f_a(1)}, f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(\cancel{4}, 8, 7, 1, 11, 2, 0, 0, 0)$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(\cancel{f_a(1)}, f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(\cancel{4}, 8, 7, 1, 11, 2, 0, 0, 0)$$

$$1 \in A_1 = \{1, 3, 9\}$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(\cancel{f_a(1)}, f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(\cancel{4}, 8, 7, 1, 11, 2, 0, 0, 0)$$

$$1 \in A_1 = \{1, 3, 9\}$$

$$\Rightarrow \delta(x) = c_3 \frac{x - 9}{3 - 9} + c_9 \frac{x - 3}{9 - 3} = 2x + 2$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(\cancel{f_a(1)}, f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(\cancel{4}, 8, 7, 1, 11, 2, 0, 0, 0)$$

$$1 \in A_1 = \{1, 3, 9\}$$

$$\Rightarrow \delta(x) = c_3 \frac{x - 9}{3 - 9} + c_9 \frac{x - 3}{9 - 3} = 2x + 2$$

$$\delta(1) = 4$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), \cancel{f_a(2)}, f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(4, 8, 7, \cancel{1}, 11, 2, 0, 0, 0)$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), \cancel{f_a(2)}, f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(4, 8, 7, \cancel{1}, 11, 2, 0, 0, 0)$$

$$2 \in A_2 = \{2, 6, 5\}$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), \cancel{f_a(2)}, f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(4, 8, 7, \cancel{1}, 11, 2, 0, 0, 0)$$

$$2 \in A_2 = \{2, 6, 5\}$$

$$\Rightarrow \delta(x) = c_6 \frac{x - 5}{6 - 5} + c_5 \frac{x - 6}{5 - 6} = 9x + 9$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), \cancel{f_a(2)}, f_a(6), f_a(5), f_a(4), f_a(12), f_a(10))$$

$$(4, 8, 7, \cancel{1}, 11, 2, 0, 0, 0)$$

$$2 \in A_2 = \{2, 6, 5\}$$

$$\Rightarrow \delta(x) = c_6 \frac{x - 5}{6 - 5} + c_5 \frac{x - 6}{5 - 6} = 9x + 9$$

$$\delta(2) = 1$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), \cancel{f_a(4)}, f_a(12), f_a(10))$$

$$(4, 8, 7, 1, 11, 2, \cancel{0}, 0, 0)$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), \cancel{f_a(4)}, f_a(12), f_a(10))$$

$$(4, 8, 7, 1, 11, 2, \cancel{0}, 0, 0)$$

$$4 \in A_3 = \{4, 12, 10\}$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), \cancel{f_a(4)}, f_a(12), f_a(10))$$

$$(4, 8, 7, 1, 11, 2, \cancel{0}, 0, 0)$$

$$4 \in A_3 = \{4, 12, 10\}$$

$$\Rightarrow \delta(x) = c_{12} \frac{x - 10}{12 - 10} + c_{10} \frac{x - 12}{10 - 12} = 0$$

$$\delta(x) = \sum_{\beta \in A_j \setminus \alpha} c_\beta \prod_{\beta' \in A_j \setminus \{\alpha, \beta\}} \frac{x - \beta'}{\beta - \beta'}$$

$$(f_a(1), f_a(3), f_a(9), f_a(2), f_a(6), f_a(5), \cancel{f_a(4)}, f_a(12), f_a(10))$$

$$(4, 8, 7, 1, 11, 2, \cancel{0}, 0, 0)$$

$$4 \in A_3 = \{4, 12, 10\}$$

$$\Rightarrow \delta(x) = c_{12} \frac{x - 10}{12 - 10} + c_{10} \frac{x - 12}{10 - 12} = 0$$

$$\delta(4) = 0$$

Example of LRC-2 code

Let $\mathbb{F} = \mathbb{F}_{13}$, $A = \mathbb{F} \setminus \{0\}$
$\mathcal{A}' = \{\{1, 5, 12, 8\}, \{2, 10, 11, 3\}, \{4, 7, 9, 6\}\}$
$\mathcal{A} = \{\{1, 3, 9\}, \{2, 6, 5\}, \{4, 12, 10\}, \{7, 8, 11\}\}$
$f_a(x) = a_0 + a_1 x + a_2 x^4 + a_3 x^6$

$$a = (1, 1, 1, 1) \quad \longrightarrow \quad c = (4, 8, 7, 5, 2, 6, 2, 2, 2, 3, 9, 1)$$

As already seen: $\delta(x) = 2x + 2$; $\delta(1) = 4$.

$$\delta'(x) = c_5 \frac{x-12}{5-12}\frac{x-8}{5-8} + c_{12}\frac{x-5}{12-5}\frac{x-8}{12-8} + c_8\frac{x-5}{8-5}\frac{x-12}{8-12}$$

$$= 6 \cdot 5 \cdot (x^2 + 6x + 5) + 2 \cdot 7 \cdot (x^2 + 1) + 9 \cdot 1 \cdot (x^2 + 9x + 8)$$

$$= x^2 + x + 2 \quad \longrightarrow \quad \delta'(1) = 4$$

Algebraic Geometric Codes

- X, Y smooth projective absolutely irreducible curves.
- $g : X \to Y$ rational separable map of degree $r + 1$
- $g* : K(Y) \to K(X)$ associated function field mapping.
- $g*$ defines a field embedding $K(Y) \hookrightarrow K(X)$ when identifying $K(Y)$ with its image.
- Primitive element theorem: $\exists x \in K(X)$ s.t. $K(X) = K(Y)(x)$ and $x^{r+1} + b_r x^r + ... + b_0 = 0$ for some $b_i \in K(Y)$.
- Denote $deg(x) = h$
- $S = \{P_1, ..., P_s\} \subset Y(K)$
- Let $D \geq 0$ a divisior, $\deg(D) = \ell \geq 1$, with $\text{supp}(D) \cap S = \emptyset$

- Assumptions:
  - $A := g^{-1}(S) = \{P_{ij}, i = 0, ..., r,\ j = 1, ..., s\} \subseteq X(K)$
  - $g(P_{ij}) = P_j$ for all i, j
  - $b_i \in L(n_i D),\ i = 0, ..., r$ for some $n_i \in \mathbb{N}$
- Let $\{f_1, ..., f_m\}$ basis of $L(D)$.
- Functions f_i contained in $K(Y) \Rightarrow$ constant on the fibers of g.
- Riemann-Roch theorem: $m \geq \ell - g_Y + 1$, (g_Y genus of Y)
- Consider the subspace
  $V := \{f_j x^i, i = 0, ..., r-1,\ j = 1, ..., m\}$
  $$\begin{array}{rccc} e := ev_A : & V & \to & K^{(r+1)s} \\ & F & \mapsto & (F(P_{ij}), i = 0, ..., r,\ j = 1, ..., s) \end{array}$$

Let $\mathcal{C}(D, g)$ be the image of the mapping: $e(V) \subseteq \mathbb{F}_q^{(r+1)s}$.

Theorem 2.1.

The subspace $\mathcal{C}(D, g)$ forms an (n, k, r) linear LRC code with the parameters:

$$\left.\begin{array}{c} n = (r+1)s \\ k = rm \geq r(\ell - g_Y + 1) \\ d \geq n - \ell(r+1) - (r-1)h \end{array}\right\}$$

Code coordinates partitioned into s subsets $A_j = \{P_{ij}, i = 0, ..., r\}$
Local recovery of erased symbol $c_{ij} = F(P_{ij})$ can be performed by polynomial interpolation through the points of A_j.

Extending this construction to:

- Families of curves (Hermitian curves and Garcia-Stichtenoth curves)
- Higher-dimensional varieties

Authors constructed optimal LRC codes with large code lentgh. E.g. $n = q^2 - 1$ and $n = q^2 + 2$.

Cyclic and Binary LRC Codes

For binary cyclic LRC codes:

- Improvement of decoding performance.
- Reduction of decoding complexity.

How:

- Ordered Statistics Decoding (OSD)
- Trellis Decoding

For linear cyclic LRC codes:

- Optimal cyclic codes that arise from RS like construction of LRC codes.
- Characterization of these codes in terms of their zeros.
- There are many equivalent ways of constructing optimal cyclic LRC codes over a given field.

List Decoding

A code $\mathcal{C}$ of length n over an alphabet A of size q is (τ, ℓ)-list decodable if:

$$\forall v \in A^n \quad |\{c \in \mathcal{C} | d(v, c) \leq \tau\}| \leq \ell$$

Johnson showed any code of length n and distance d is (τ_J, ℓ)-list decodable, where:

- $\tau_J = n - \sqrt{n(n-d)}$ the Johnson radius
- Size of list ℓ polynomial in n.

List Decoding

Holzbaur and Wachter-Zeh improved the Johnson radius for LRC codes

- List decode each recovering set
- List decode the LRC code using combinations of previously decoded recovering sets.

Complexity and list size polynomial in n when number of recovering sets remain constant.

Work Plan

Open Problems

Improvement of Bounds

Bound appears to be far from tight.

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

Authors believe largest possible rate for (n, k, r, t)-LRC code is $\left(\frac{r}{r+1}\right)^t$ as long as t not too large (e.g. $t \in O(\log n)$).

Algebraic Geometric Codes

Search for new constructions of LRC codes over algebraic varieties.
Looking for optimal LRC codes with small field size.

Binary and Cyclic LRC codes

Improve decoding of binary LRC codes, non cyclic.

List Decoding

Search for new families of LRC codes that could be list decoded beyond Johnson radius.

Time Plan and Methodology

First Year

Goal: Research of the state of the art, and stating the problems to study.

- Master's Degree in Advanced Mathematics and Mathematical Engineering
- Visitor student in University of Maryland, with prof. Alexander Barg
- Algebraic Geometry Seminar. Two sessions per week.
  - *"Algebraic Curves"*. William Fulton
  - Lecture notes on Algebraic Geometry, Andreas Gathmann
  - Lecture notes on Plane Algebraic Curves, Andreas Gathmann
- Self-Learning: *"The Probabilistic Method"*, Alon and Spencer.

Second Year

Goal: Work on the stated problems.

- Seminars
  - Algebraic Geometric Codes, two sessions per week.
    - *"Algebraic Geometric Codes: basic notions"*; Vladut, Nogin and Tsfasman.
    - *"Algebraic Function Fields and Codes"*, Stichtenoth.
  - Probabilistic Method, two sessions per week.
    - *"The Probabilistic Method"*, Alon and Spencer.
- Self-Learning: *"The Theory of Error-Correcting Codes"*, MacWilliams and Sloane.

Third Year

Goal: Conclude research and write thesis.

- Conclusion of the research
- Writing of PhD thesis.