IERG6120 Coding for Distributed Storage Systems Lecture 9 - 18/10/2016

Non-linear LRC and the bound by Gopalan et al.

Lecturer: Kenneth Shum *Scribe: Cao Qi*

Notations: Let $C \subseteq \mathbb{F}_q^n$ be a code, $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ denote a codeword in C. For any subset $I = \{i_1, i_2, \dots, i_m\}$ of the index set $[n] := \{1, 2, \dots, n\}$, let

$$C_I := \{(x_{i_1}, x_{i_2}, \dots, x_{i_m}) : \boldsymbol{x} = (x_1, x_2, \dots, x_n) =\in C\}$$

be the *restriction* of C to I.

We note that $|C_I| \leq |C_J|$ whenever $I \subseteq J \subseteq [n]$.

Definition 1. *For $i \in [n]$, we say that code symbol i has* locality r *if there exists an index set $I \subseteq [n] \setminus \{i\}$ such that $|I| \leq r$ and $|C_I| = |C_{I\cup\{i\}}|$. A code is said to have* all-symbol locality *if for any $i \in \{n\}$ there exists an $I \subseteq [n] \setminus \{i\}$ such that $|I| \leq r$ and $|C_I| = |C_{I\cup\{i\}}|$. For a systematic code, if these properties apply to its systematic symbols, then the code is said to have* information locality r*.*

Example 1 The generator matrix is shown as following.

$$G = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

We can obtain the corresponding code:

$$C = \{0000, 1011, 1101, 0110\}.$$

When $i = 1$, we can find an $I = \{4\}$ such that $C_{\{4\}} = \{0, 1\}$ and $C_{\{1,4\}} = \{00, 11\}$, i.e., $|C_I| = |C_{I\cup\{i\}}|$. Thus symbol 1 has locality 1.

When $i = 2$, we can find an $I = \{1, 3\}$ such that $C_{\{1,3\}} = \{00, 01, 10, 11\}$ and $C_{\{1,2,3\}} = \{000, 101, 110, 011\}$. Thus symbol 2 has locality 2.

Notation 1. *Let $(n, k, r)_q$-LRC, which is an abbreviation of Locally Repairable Code, denote a code of length n, containing q^k codewords and having all-symbol locality r.*

Theorem 2. *If C is an $(n, k, r)_q$-LRC, then we have*

- $\frac{k}{n} \leq \frac{r}{r+1}$,
- $d(C) \leq n - k - \lceil \frac{k}{r} \rceil + 2$.

In particular when $k = r$, we have $d(C) \leq n - k + 1$, which is the Singleton bound.

The inequality in Theorem 2 was first proved by Gopalan *et al.* for codes with information locality [1]. In this notes, we follows the proof in [2], which is for codes with all-symbol locality.

Lemma 3. *In a nonlinear q-ary code of length n, the minimum distance is characterized by*

$$d = n - \max_{I\subseteq[n]}\{|I| : |C_I| < |C|\}.$$

Proof Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be distinct codewords in C. If $d(\boldsymbol{u},\boldsymbol{v}) = d'$, then we can find an index set I of size $n-d'$, so that the two codewords are identical precisely at the positions indexed by I, and $|C_I| < |C|$. Since $\boldsymbol{u}$ and $\boldsymbol{v}$ are assumed to be distinct, the index set I is not equal to $[n]$. We obtain

$$\begin{aligned} d = \min_{\substack{\boldsymbol{u},\boldsymbol{v}\in C\\ \boldsymbol{u}\neq\boldsymbol{v}}} d(\boldsymbol{u},\boldsymbol{v}) &= \min_{I\subsetneq[n]} \{n-|I| : |C_I| < |C|\} \\ &= \min_{I\subseteq[n]} \{n-|I| : |C_I| < |C|\} \\ &= n - \max_{I\subseteq[n]} \{|I| : |C_I| < |C|\}. \end{aligned}$$

□

Consider a directed graph $G = (V,E)$, where $V = [n]$ and $E \subseteq V \times V$. Suppose that the out degrees of the vertices are $d_1, d_2, \ldots, d_n$. Let $G_U = (U, E_U)$ be the induced graph on a vertex subset U, where $U \subseteq V$ and $E_U := \{(a,b) \in E : a,b \in U\}$.

Theorem 4. *There exists an induced subgraph that is acyclic, with at least* $\frac{n}{1+\frac{1}{n}\sum_{i=1}^{n} d_i}$ *vertices.*

Proof Pick a random permutation $\pi : [n] \to [n]$. Let U_π be a subset of V, defined by $i \in U_\pi$ iff for every outgoing edge (i,j), $\pi(i) < \pi(j)$. Check that induced graph of U_π has no cycle. We can define a function on i as following.

$$\mathbb{1}_i = \begin{cases} 1, & i \in U_\pi \\ 0, & i \notin U_\pi. \end{cases}$$

Then we can obtain the expect value of the number of vertices in E_U,

$$\begin{aligned} E(U_\pi) &= E(\mathbb{1}_1) + E(\mathbb{1}_2) + \cdots + E(\mathbb{1}_n) \\ &= \frac{1}{1+d_1} + \frac{1}{1+d_2} + \cdots + \frac{1}{1+d_n} \\ &\overset{(a)}{\geq} \frac{n}{1+\frac{1}{n}\sum_{i=1}^{n} d_i}, \end{aligned}$$

where (a) follows from that $A.M. \geq H.M.$ i.e., the arithmetic mean is larger than or equal to harmonic mean. □

Proof (of Theorem 2)

1) Firstly, we consider the number of the redundant symbols. For each $i \in [n]$, there exists an index set $I_i \subseteq [n] \setminus \{i\}$ of size less than or equal to r, such that symbol i can be repaired by symbols indexed by I_i. There may be more than one choice of such I_i, and we only need to pick one for each i.

We construct a directed graph $G = (V,E)$ on n vertices. We label the vertices from 1 to n. For $i \in [n]$, we draw an edge from node i to node j for each $j \in I_i$. The out-degree of node i is $|I_i|$.

By Theorem 4, there exists a subset $U \subseteq V$ of size

$$|U| \geq \frac{n}{1+\frac{1}{n}\sum_{i=1}^{n} |I_i|},$$

such that the induced graph G_U containing no directed cycle. Considering that $|I_i| \leq r$, we can obtain that

$$|U| \geq \frac{n}{1+\frac{1}{n}\sum_{i=1}^{n} d_i} \geq \frac{n}{1+r}.$$

If vertex $i \in U$ has no out-going edge in E_U, then $I_i \subset U^c$, i.e., symbol i is a function of the code symbols indexed by U^c. (The notation U^c signifies the complement of U in V.) Repeat the argument for $G_{U\setminus\{i\}}$, we will eventually get an empty graph. Therefore, each code symbol in U is a function of code symbol in U^c. In other words, the code symbols with indices in U are redundant symbols. The number of codewords q^k must be less than or equal to $q^{n-|U|} \leq q^{nr/(1+r)}$.

Hence,

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

and this implies the first part of Theorem 2.

2) Next, we try to find an index set I such that $|C_I| < q^k$. Since

$$|U| \geq \frac{n}{1+r} \geq \frac{k}{r} \geq \lfloor \frac{k-1}{r} \rfloor,$$

we can pick a subset of $U' \subseteq U$ of size $\lfloor \frac{k-1}{r} \rfloor$. The choice of U' is arbitrary as long as the size of U' is equal to $\lfloor \frac{k-1}{r} \rfloor$.

Let $\mathcal{N} := \left(\bigcup_{i\in U'} I_i\right) \setminus U'$ be the neighborhood of U'. The symbols in U' are uniquely determined by symbols in $\mathcal{N}$. Considering that $|I_i| \leq r$, we have

$$|\mathcal{N}| \leq r|U'| = r\lfloor \frac{k-1}{r} \rfloor \leq k-1.$$

As $|\mathcal{N}| \leq k-1$, we can pick a set in $(U')^c$ with size $k-1$ exactly, denoted by $\mathcal{N}'$, such that $\mathcal{N} \subseteq \mathcal{N}'$. If $|\mathcal{N}| = k-1$ then this step is trivial, otherwise we can arbitrarily pick any $k-1-|\mathcal{N}|$ elements in $(U')^c$ and add them to $\mathcal{N}$. This can always be done because

$$\begin{aligned} n - |U'| &= n - \lfloor \frac{k-1}{r} \rfloor \\ &\geq k\frac{r+1}{r} - \frac{k-1}{r} \\ &\geq k + \frac{k}{r} - \frac{k}{r} + \frac{1}{r} \\ &> k > k-1. \end{aligned}$$

Thus, the symbols in U' are determined by $\mathcal{N}'$, and

$$|C_{U'\cup\mathcal{N}'}| = |C_{\mathcal{N}'}| \leq q^{k-1}.$$

The last inequality follows from $|\mathcal{N}'| = k-1$.

We have already found an index set $I = U' \cup \mathcal{N}'$ such that $|C_I| < q^k$, so we have

$$\max_I \{|I| : |C_I| < q^k\} \geq |U' \cup \mathcal{N}'| = k - 1 + \lfloor \frac{k-1}{r} \rfloor.$$

Thus we can obtain

$$d = n - \max_{I\subseteq[n]} \{|I| : |C_I| < q^k\} \leq n - (k - 1 + \lfloor \frac{k-1}{r} \rfloor) \overset{(a)}{=} n - k - \lceil \frac{k}{r} \rceil + 2,$$

where (a) follows from that $1 + \lfloor \frac{k-1}{r} \rfloor = \lceil \frac{k}{r} \rceil$. $\square$

Exercises

1. [3] We say that a code symbol with index i has t *disjoint repair sets* if we can find t disjoint index sets I_i^j, for $j = 1, 2, \dots, t$, such that (i) $I_i^j \subset [n] \setminus \{i\}$ for all j, (ii) $I_i^j \cap I_i^\ell = \emptyset$ for all $j \neq \ell$, and (iii) for each $j = 1, 2, \dots, t$, we can recover the code symbol at location i from the code symbols indexed by I_i^j. For an (n, k, r)-LRC C in which all code symbols have t distinct repair sets of size less than or equal to r, prove that
$$\frac{k}{n} \leq \prod_{j=1}^{t} \frac{1}{1 + \frac{1}{jr}},$$
and
$$d(C) \leq n - \sum_{j=0}^{t} \Big\lfloor \frac{k-1}{r^j} \Big\rfloor.$$

2. This exercise is the analog of Theorem 4 for undirected graph. For an undirected graph G on vertex set V, a subset U of the vertex set is called an *independent set* if no two vertices in U are adjacent in G. The size of the largest independent set of an undirected graph G is called the *independence number of* G, and is commonly denoted by $\alpha(G)$.

   If G is an directed graph in which the vertices has maximal degree D, then it is easy to show that
$$\alpha(G) \geq \frac{n}{1+D}$$
   where n is the number of vertices. Indeed, we can iteratively create an independence set. The procedure is: (i) arbitrarily select a vertex v in G, (ii) remove v and its adjacent vertices from G, repeat (i) and (ii) until we obtain an empty graph. In each iteration we remove at most
$$1 + D$$
   vertices, and hence at least $n/(1 + D)$ vertices are selected in the process.

   If the degrees of all vertices are known, then we can have a better bound. Suppose that the degrees of the n vertices in graph G are $d_1, d_2, \dots, d_n$, which may or may not be equal to each other. Prove that
$$\alpha(G) \geq \sum_{i=1}^{n} \frac{1}{1 + d_i} \geq \frac{n}{1 + (d_1 + d_2 + \ldots + d_n)/n}.$$
   See the "The probabilistic lens: Turán's theorem" in [4].

References

[1] 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.

[2] 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.

[3] I. Tamo, A. Barg and A. Frolov, "Bounds on the parameters of locally repairable codes," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp.3070–3083, Jun. 2016.

[4] N. Alon and J. H. Spencer, *The probabilistic method*, 2nd edition, John Wiley & Son, New York, 2004.