Linear Locally Decodable Codes **Fall 2016**

Lecture 8: Locally Correctable Codes

Lecturer: Zeev Dvir *Scribe: Kalina Petrova*

Notice that the Hadamard code and the Reed-Muller code have the following stronger properties than simply what is characteristic for Locally Decodable Codes.

- The message is part of the codeword, which is to say that the code is *systematic*. This means that if the code is $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$, then the identity matrix of size k is part of the generating matrix.
- The decoder can locally decode not only any symbol from the message, but also any symbol from the codeword.

Example 8.1. For instance, the Hadamard code $E : \mathbb{F}_2^k \to \mathbb{F}_2^n$, where $n = 2^k$, is given by $E(\mathbf{x}) = \Big(\langle \mathbf{x}, \mathbf{a} \rangle\Big)_{\mathbf{a} \in \mathbb{F}_2^k}$. To decode some element of $E(\mathbf{x})$ indexed by some $\mathbf{a} \in \mathbb{F}_2^k$, we need to find $\langle \mathbf{x}, \mathbf{a} \rangle$, but since this element of the codeword might have an error, we pick a random $\mathbf{b} \in \mathbb{F}_2^k$, and we calculate what we're looking for using the formula $\langle \mathbf{x}, \mathbf{a} \rangle = \langle \mathbf{x}, \mathbf{b} \rangle + \langle \mathbf{x}, \mathbf{b}+\mathbf{a} \rangle$. The probability of error is the same as the probability of error when calculating some element of the original message since the method of calculation is the same.

These properties are at the heart of the concept *Locally Correctable Codes*.

Definition 8.1. An r-query Linear Locally Correctable Code with error δ is a linear map $E : \mathbb{F}_p^k \to \mathbb{F}_p^n$ such that there is a decoder $D(i, \mathbf{y})$, $i \in [n]$, that queries at most r positions in $\mathbf{y}$ and, if $dist(\mathbf{y}, E(\mathbf{x})) \leq \delta$, returns $E(\mathbf{x})_i$ with probability at least $\frac{3}{4}$.

By the same argument as for Theorem 1.1 for LDC's, we can show the following theorem.

Theorem 8.1 (Structure Theorem)**.** Let $E : \mathbb{F}_p^k \to \mathbb{F}_p^n$ be an r-query Locally Correctable Code with error δ. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}_p^k$ be the rows of the generating matrix of E, which is to say that $E(\mathbf{x})_i = \langle \mathbf{v}_i, \mathbf{x} \rangle$. Then $\forall i \in [n]$, there exists an r-matching $M^i = (T_1^i, \ldots, T_{m_i}^i)$ on $[n]$ such that $m_i \geq \frac{\delta n}{r}$ and $\forall T_j^i \in M^i$, we have $\mathbf{v}_i \in \text{span}\{\mathbf{v}_j | j \in T_j^i\}$.

Note that the statement of Theorem 8.1 is invariant under change of generating matrix. This means that being a Locally Correctable Code is a property of the subspace $V = Im(E) \subseteq \mathbb{F}_p^n$, and we are interested in $k = \dim(V)$ as a function of n.

From now on, we will specify a Locally Correctable Code E using a list $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \in \mathbb{F}_p^k$ – the rows of the generating matrix of E, and n r-matchings $\{M^1, \ldots, M^n\}$ as above, with $\forall i \in [n], |M^i| \geq \frac{\delta n}{r} = \Omega(n)$.

Example 8.2. The Hadamard code is specified by a list $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \mathbb{F}_2^k$ and 2-matchings $\{M_{\mathbf{b}} = \{(\mathbf{a}, \mathbf{a}+\mathbf{b})\}_{\mathbf{a}\in\mathbb{F}_2^k} | \forall \mathbf{b} \in \mathbb{F}_2^k\}$. We have that $|M_{\mathbf{b}}| = 2^{k-1} = \frac{n}{2}$.

Exercise 8.1. Determine the set of matchings that define the Low-Degree Extension code/Reed-Muller code as a Locally Correctable Code.

Exercise 8.2. Show that from any Locally Correctable Code we can obtain a Locally Decodable Code with the same parameters.

The main question we are interested in is whether LCC's are actually stronger than LDC's, that is, whether there are LDC's from which we cannot obtain LCC's with the same parameters.

Exercise 8.3. Identify why the construction of Matching Vector codes does not yield a Locally Correctable Code in any obvious way.

For this reason it is believed that Matching Vector codes are not LCC's.

Currently, the best (and only) r-query Locally Correctable Codes for any r are those coming from Reed-Muller Codes. They have the form $RM_{d,t} : \mathbb{F}_q^{\binom{t+d}{t}} \to \mathbb{F}_q^{q^t}$ and $d+1$ queries, so $n = q^t \geq 2^t$ and $k = \binom{t+d}{d} \leq (t+d)^d \leq t^{r-1}$, so we have $t \geq k^{\frac{1}{r-1}}$ and thus $n \geq 2^t \geq 2^{k^{\frac{1}{r-1}}}$. It is conjectured that this is optimal.

Conjecture 8.1. There is no Locally Correctable code with $n < 2^{k^{\frac{1}{r-1}}}$.

Choice of ground field

Notice that the Hadamard code can be extended over any field $\mathbb{F}_p$ as a 2-LDC. We have $E : \mathbb{F}_p^k \to \mathbb{F}_p^n$, where $n = 2^k$, $E(\mathbf{x}) = \big(\langle \mathbf{a}, \mathbf{x}\rangle\big)_{\mathbf{a}\in\{0,1\}^k}$. The rows of the generating matrix are $\{0,1\}^k \subseteq \mathbb{F}_p^k$. Indeed, we can locally decode any $\mathbf{x}_i$ by picking a random $\mathbf{b} \in \{0,1\}^k$ and if $\mathbf{b}_i = 1$, taking $\mathbf{x}_i = \langle \mathbf{b}, \mathbf{x}\rangle - \langle \mathbf{b}^{\wedge i}, \mathbf{x}\rangle$, where $\mathbf{b}^{\wedge i}$ is $\mathbf{b}$ with the i-th bit flipped, and if $\mathbf{b}_i = 0$, taking $\mathbf{x}_i = \langle \mathbf{b}^{\wedge i}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$. However, this does not work in general if we want to claim that E is an LCC, because then we would need to be able to obtain $\langle \mathbf{a}, \mathbf{x}\rangle$ for any $\mathbf{a} \in \{0,1\}^k$ as a combination of two other entries of $E(\mathbf{x})$, and $\langle \mathbf{a}, \mathbf{x}\rangle = \langle \mathbf{b}, \mathbf{x}\rangle + \langle \mathbf{b}+\mathbf{a}, \mathbf{x}\rangle$ does not hold for any $\mathbb{F}_p^k$, because for each i such that $\mathbf{a}_i = 0$ and $\mathbf{b}_i = 1$, the contribution of the i-th coordinate to $\langle \mathbf{b}, \mathbf{x}\rangle + \langle \mathbf{b}+\mathbf{a}, \mathbf{x}\rangle$ is $2\mathbf{x}_i$ and the contribution of the i-th coordinate to $\langle \mathbf{a}, \mathbf{x}\rangle$ is 0. A natural question to ask is whether we can get a 2-LCC analogous to Hadamard code, for instance with $n = 2^k$, over larger fields. The answer is yes: we can take $\mathbb{F}_q$ with $q = 2^l$. Since $\mathbb{F}_2$ is a subfield of $\mathbb{F}_q$, $char(\mathbb{F}_q) = 2$, because any subfield of a field has the same characteristic as the field does. Then we have $E : \mathbb{F}_{2^l}^k \to \mathbb{F}_{2^l}^n$, where $n = 2^k$, and $E(\mathbf{x}) = \big(\langle \mathbf{a}, \mathbf{x}\rangle\big)_{\mathbf{a}\in\{0,1\}^k}$. Notice that $\forall \mathbf{a}, \mathbf{b} \in \{0,1\}^k$, $\forall i \in [k]$, $\mathbf{x}_i\mathbf{a}_i = \mathbf{x}_i\mathbf{b}_i + \mathbf{x}_i\big(\mathbf{a}+\mathbf{b}\big)_i$, since $\mathbf{x}_i + \mathbf{x}_i = 0$ in $\mathbb{F}_q$ because $char(\mathbb{F}_q) = 2$. Therefore, $\forall \mathbf{a}, \mathbf{b} \in \{0,1\}^k$, $\langle \mathbf{x}, \mathbf{a}\rangle = \langle \mathbf{x}, \mathbf{b}\rangle + \langle \mathbf{x}, \mathbf{a}+\mathbf{b}\rangle$, so any position of $E(\mathbf{x})$ can be decoded as before.

Is there an LCC over $\mathbb{F}_p$, where $p > 2$ is prime? The only known such 2-LCC has $n = p^k$ and $E(\mathbf{x}) = \big(\langle \mathbf{x}, \mathbf{a}\rangle\big)_{\mathbf{a}\in\mathbb{F}_p^k}$. Then if we want to decode $E(\mathbf{x})_{\mathbf{a}}$ for some $\mathbf{a} \in \mathbb{F}_p^k$, we can take a random $\mathbf{b} \in \mathbb{F}_p^k$ and notice that $\langle \mathbf{x}, \mathbf{a}\rangle = \langle \mathbf{x}, \mathbf{b}\rangle + \langle \mathbf{x}, \mathbf{a} - \mathbf{b}\rangle$.

Theorem 8.2. [BDSS11] A 2-LCC over $\mathbb{F}_p^k$, where p is prime, has to have $n \geq C_{p,\delta} p^{\Omega(\delta k)}$.

If we fix p, we see that this theorem implies n grows at least exponentially in k.

Proof. It is easier to think about bounding $k = \dim(\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\})$ as a function of n. It is enough to show that $k \leq C_{p,\delta} + O(\frac{1}{\delta}\log_p n)$, where $C_{p,\delta}$ is some constant that depends only on p and δ. We will first show a lemma which we will then use to prove this.

Lemma 8.1. Let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subseteq \mathbb{F}_p^k$ be a 2-LCC with error δ. Suppose there are no repetitions in V, that is, $\forall i, j \in [n]$ such that $i \neq j$, $\forall \alpha \in \mathbb{R}$, $\mathbf{v}_i \neq \alpha \mathbf{v}_j$. Then $\exists V' \subseteq V$ with $|V'| \geq \big(\frac{\delta}{p}\big)^c n$, where c is some absolute constant, such that $\dim(V') = \dim(\text{span}\{\mathbf{v}_j | \forall \mathbf{v}_j \in V'\}) \leq C_{p,\delta} + \log_p n$.

Proof. The proof uses two tools from additive combinatorics. The first one is the Balog-Szemeredi-Gowers Lemma.

Lemma 8.2. [BS94, Gow98] Let $A \subseteq G$, where G is any finite abelian group (say $G = \mathbb{F}_p^k$). Suppose that $|\{(a_1, a_2) \in A \times A | a_1 + a_2 \in A\}| \geq \alpha |A|^2$. Then $\exists A' \subseteq A$ such that $|A'| \geq \alpha^c |A|$ and $|A' + A'| \leq \alpha^{-c} |A'|$, where c is an absolute constant and $A' + A' = \{a_1 + a_2 | a_1, a_2 \in A'\}$ is called the *sum-set* of A'.

Intuitively, the lemma above states that if A does not grow much when added to itself, then there exists a large $A' \subseteq A$ with small sum-set.

The second tool we use is Ruzsa's theorem.

Theorem 8.3. [GR07]

Let $A \subseteq \mathbb{F}_p^k$ be such that $|A + A| \leq \mu |A|$, then there exists a subspace W of $\mathbb{F}_p^k$ such that

1. $A \subseteq W$
2. $|W| \leq \mu^c p^{\mu^c} |A|$, where c is an absolute constant.

These two statements imply that $\dim(A) \leq \log_p |W| \leq C_{p,\mu} + \log_p |A|$, where $C_{p,\mu}$ is a constant that depends only on p and μ.

Here we continue the proof of Lemma 8.1. We are given the 2-LCC $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subseteq \mathbb{F}_p^k$. Let $U \subseteq \mathbb{F}_p^k$ be defined as $U = \{a\mathbf{v}_i | a \in \mathbb{F}_p, a \neq 0, \mathbf{v}_i \in V\}$. Since our 2-LCC has no repetitions, we have $|U| = (p-1)n$, as $\mathbf{v}_1, 2\mathbf{v}_1, \ldots, (p-1)\mathbf{v}_1, \mathbf{v}_2, 2\mathbf{v}_2, \ldots, (p-1)\mathbf{v}_2, \ldots, \mathbf{v}_n, 2\mathbf{v}_n, \ldots, (p-1)\mathbf{v}_n$ are all different. We use the following claim.

Claim. $|\{(u,u') \in U \times U | u+u' \in U\}| \geq \Omega(\frac{\delta}{p})|U|^2$.

Proof. For any $u = a\mathbf{v}_i \in U$, consider the matching M^i from the 2-LCC. We have that $|M^i| \geq \frac{\delta n}{2}$. For some pair $(\mathbf{v}_j, \mathbf{v}_k) \in M^i$, suppose $\mathbf{v}_i = b\mathbf{v}_j + c\mathbf{v}_k$. Then we have that $\forall a \in \mathbb{F}_p, a \neq 0$, $a\mathbf{v}_i + (-ab)\mathbf{v}_j = (ac)\mathbf{v}_k \in U$, so $(a\mathbf{v}_i, (-ab)\mathbf{v}_j) \in \{(u,u') \in U \times U | u+u' \in U\}$. Thus for each element of U $a\mathbf{v}_i$, we get that it participates in $\Omega(\delta n)$ pairs like this, so there are a total of $\Omega(\delta npn) = \Omega(\frac{\delta}{p})|U|^2$ pairs in U with sum inside U. Notice that it is possible to have counted some pairs twice by changing the order of the two elements of U within the pair, but this does not change the order of magnitude, and we have not counted any pair more than twice.

□

Using this claim, we finish the proof of Lemma 8.1. By the Balog-Szemeredi-Gowers Lemma, since $|\{(u,u') \in U \times U | u+u' \in U\}| \geq \Omega(\frac{\delta}{p})|U|^2$, then $\exists U' \subseteq U$ such that $|U'| \geq (\frac{\delta}{p})^c|U|$ and $|U'+U'| \leq (\frac{\delta}{p})^{-c}|U'|$. Then applying Ruzsa's Theorem by setting $A = U'$ and $\mu = (\frac{\delta}{p})^{-c}$, we get that $\dim(U') \leq C_{p,\delta} + \log_p |U'|$. Now we can modify U' to obtain $U'' \subseteq V$ by doing the following. For each $a\mathbf{v}_i \in U'$, where $\mathbf{v}_i \in V$ and $a \in \mathbb{F}_p$, take $\mathbf{v}_i \in U''$. In this way we lose at most a factor of p in the size, that is, $|U''| \geq \frac{1}{p}|U'| \geq (\frac{\delta}{p})^{c'}|U| \geq (\frac{\delta}{p})^{c''}n$ for some constants c' and c'' since $|U| = pn$, and it still holds that $\dim(U'') \leq C_{p,\delta} + \log_p |U'| \leq C_{p,\delta} + \log_p (pn) \leq C'_{p,\delta} + \log_p n$. This completes the proof of Lemma 8.1.

□

To finish the proof of Theorem 8.2, we amplify V' (we expand it until it turns into V) without increasing its dimension too much. This is done in two steps, but we do not discuss the details here since they are highly technical. We refer the interested reader to the original paper [BDSS11].

□

Exercise 8.4. Show that replacing A with A' is necessary in Lemma 8.2. That is, give an A with $|\{a \in A | a+a \in A\}| \geq \frac{|A|^2}{k}$ but such that $A+A$ is very large.

References

[BDSS11] Arnab Bhattacharyya, Zeev Dvir, Shubhangi Saraf, and Amir Shpilka. Tight lower bounds for 2-query LCCs over finite fields. *Proc. of FOCS 2011*, page 638–647, 2011.

[BS94] Antal Balog and Endre Szemeredi. A statistical theorem of set addition. *Combinatorica*, 14:263–268, 1994.

[Gow98] William Timothy Gowers. A new proof of szemeredi's theorem for progressions of length four. *GAFA*, 8:529–551, 1998.

[GR07] Ben Green and Imre Z. Ruzsa. Freiman's theorem in an arbitrary abelian group. *Journal of the London Mathematical Society*, 75:163–175, 2007.