Linear Locally Decodable Codes **Fall 2016**

Lecture 1: Introduction

Lecturer: Zeev Dvir *Scribe: Kalina Petrova*

Preliminaries

We start with some notation that will be used throughout these notes.

- $[k] = \{1, 2, \ldots, k\}$;
- $\mathbb{F}_q$ is a finite field of q elements;
- $\forall \mathbf{x}, \mathbf{y} \in \mathbb{F}_q^n$, we will use $dist(\mathbf{x}, \mathbf{y})$ to denote the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$, that is, the number of positions in which $\mathbf{x}$ and $\mathbf{y}$ differ.
- We will use $\log x$ to denote $\log_2 x$ for any x.

Definition 1.1. A linear map $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is an (r, δ, ε)-LDC (Locally Decodable Code) if there exists a randomized procedure D such that

1. For each $\mathbf{x} \in \mathbb{F}_q^k$ and $\forall \mathbf{y} \in \mathbb{F}_q^n$ s.t. $dist(E(\mathbf{x}), \mathbf{y}) \leq \delta n$, we have $\forall i \in [k], Pr[D(\mathbf{y}, i) = \mathbf{x}_i] \geq 1 - \varepsilon > \frac{1}{q}$.
2. D makes at most r non-adaptive queries to $\mathbf{y}$.

In other words, a linear map E is a Locally Decodable Code if each coordinate of the original message can be decoded with high probability querying only r coordinates of the encoded message. The reason why we need this high probability to be higher than $\frac{1}{q}$ is that a procedure which makes a uniformly random guess for the coordinate in question will have a success probability $\frac{1}{q}$, since there are q symbols in use.

Sometimes we will omit δ, ε, saying E is an r-LDC if it is an (r, δ, ε)-LDC for some $\delta, \varepsilon > 0$.

Note that if we can afford to increase $O(r)$ by multiplying it with a constant, then we can fix $\varepsilon = \frac{1}{4}$. This is because an LDC with error $\frac{1}{4}$ can always be amplified to obtain an LDC with error ε', where ε' can be arbitrarily small. We can do this by running the decoding algorithm repeatedly and then taking the majority answer. Thus, by modifying r by a constant, we can get an *LDC* with any error.

Example 1.1. Hadamard code

In the case of Hadamard code, for any field $\mathbb{F}_q$, we have $n = 2^k, r = 2$. The coordinates of $E(\mathbf{x})$ are indexed by elements of $\{0,1\}^k$, and if $E_{\mathbf{v}}(\mathbf{x})$ denotes the $\mathbf{v}$-th coordinate of $E(\mathbf{x})$, $\mathbf{v} \in \{0,1\}^k \subseteq \mathbb{F}_q^k$, then $E_{\mathbf{v}}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{v} \rangle$, that is, the dot product of $\mathbf{x}$ and $\mathbf{v}$, which is defined as $\langle \mathbf{x}, \mathbf{v} \rangle = \sum_{1 \leq i \leq k} \mathbf{v}_i \mathbf{x}_i$.

The local decoder $D(\mathbf{y}, i)$ operates in the following manner. Pick a $\mathbf{v} \in \{0,1\}^k$ uniformly at random. Let $\tilde{\mathbf{v}}$ be obtained from $\mathbf{v}$ by flipping the i-th bit. Then our guess for $\mathbf{x}_i$ is

$$D(\mathbf{y}, i) = \begin{cases} \mathbf{y}_{\mathbf{v}} - \mathbf{y}_{\tilde{\mathbf{v}}}, & \text{if } \mathbf{v}_i = 1 \\ \mathbf{y}_{\tilde{\mathbf{v}}} - \mathbf{y}_{\mathbf{v}}, & \text{if } \mathbf{v}_i = 0 \end{cases}$$

If both $\mathbf{y}_{\mathbf{v}} = E_{\mathbf{v}}(\mathbf{x})$ and $\mathbf{y}_{\tilde{\mathbf{v}}} = E_{\tilde{\mathbf{v}}}(\mathbf{x})$, then we will get the right answer, because:

- If $\mathbf{v}_i = 1$, then $\mathbf{y}_{\mathbf{v}} - \mathbf{y}_{\tilde{\mathbf{v}}} = E_{\mathbf{v}}(\mathbf{x}) - E_{\tilde{\mathbf{v}}}(\mathbf{x}) = \sum_{1 \le j \le k, \mathbf{v}_j = 1} \mathbf{x}_j - \sum_{1 \le j \le k, \tilde{\mathbf{v}}_j = 1} \mathbf{x}_j = \mathbf{x}_i$, since $\mathbf{v}$ and $\tilde{\mathbf{v}}$ only differ in the i-th bit.
- If $\mathbf{v}_i = 0$, then $\mathbf{y}_{\tilde{\mathbf{v}}} - \mathbf{y}_{\mathbf{v}} = E_{\tilde{\mathbf{v}}}(\mathbf{x}) - E_{\mathbf{v}}(\mathbf{x}) = \sum_{1 \le j \le k, \tilde{\mathbf{v}}_j = 1} \mathbf{x}_j - \sum_{1 \le j \le k, \mathbf{v}_j = 1} \mathbf{x}_j = \mathbf{x}_i$, since $\mathbf{v}$ and $\tilde{\mathbf{v}}$ only differ in the i-th bit.

Now since $\mathbf{v}$ and $\tilde{\mathbf{v}}$ are both chosen uniformly at random (although not independently), the probability that $\mathbf{y}_{\mathbf{v}} \neq E_{\mathbf{v}}(\mathbf{x})$ is δ, and the probability that $\mathbf{y}_{\tilde{\mathbf{v}}} \neq E_{\tilde{\mathbf{v}}}(\mathbf{x})$ is also δ, and thus the probability that at least one of these happens is no more than 2δ by the Union Bound ($Pr[\cup_i A_i] \le \sum_i Pr[A_i]$, where $A_1, A_2, A_3, \ldots$ is a countable set of events). Thus $Pr[D(\mathbf{y}, i) = \mathbf{x}_i] \ge 1 - 2\delta$, so we can set $\varepsilon = 2\delta$, and this works for any $\delta < \frac{1}{2}$.

Notice that for each coordinate i of the original message, there is a perfect matching on pairs of codeword coordinates, among which the decoder picks one pair uniformly at random. We will later prove such a matching structure always exists.

Observe that the Hadamard code is a *systematic code*, which means that the message is part of the encoding ($\mathbf{x}_i$ is at position $\mathbf{v}^i$ in the message, where $\mathbf{v}^i$ is such that its only non-zero bit is the i-th bit).

Any linear LDC $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is non-degenerate, which means that $Im(E) \subseteq \mathbb{F}_q^n$ is k-dimensional (otherwise there would be collisions, that is, there would exist $\mathbf{x} \neq \mathbf{x'}$ with $E(\mathbf{x}) = E(\mathbf{x'})$, and we wouldn't be able to recover each bit of the original message with high probability).

Definition 1.2. The rate of an LDC is $rate(E) = \frac{k}{n}$, since for any k symbols of information, the code generates n symbols.

There are two ways to define a linear LDC:

1. $E(\mathbf{x}) = A\mathbf{x}$, where A is an $n \times k$ matrix, called the generating matrix. For instance, the generating matrix of the Hadamard code is a $2^k \times k$ matrix whose rows are all binary vectors.
2. Using a parity-check $(n-k) \times n$ matrix T such that $Im(E) = Ker(T)$, that is, $\mathbf{y} \in Im(E)$ if and only if $T\mathbf{y} = \mathbf{0}$. Note that this representation is not unique (we could perform any sequence of invertible row-operations on the parity-check matrix and its kernel will be the same).

Exercise 1.1. Define the Hadamard code using a parity-check matrix.

Definition 1.3. The minimum distance of E is defined as the minimum Hamming distance between any two distinct codewords of E, $Min\_dist(E) = \min_{\mathbf{y},\mathbf{y'} \in Im(E), \mathbf{y} \neq \mathbf{y'}} \{dist(\mathbf{y}, \mathbf{y'})\}$.

Exercise 1.2. Show that if E is decodable (locally or not) from δn errors (adversarial, that is, we can't assume the errors are in random positions), then $Min\_dist(E) > 2\delta n$.

Exercise 1.3. Show that if $Min\_dist(E) > 2\delta n$, then E is decodable from δn adversarial errors (where we allow the decoder to run in exponential time).

Structure of linear LDCs

We will prove an important property of the structure of linear LDCs, which will be useful for proving lower bounds.

Definition 1.4. An r-matching on $[n]$ is a family of disjoint r-tuples $M = (T_1, T_2, \ldots, T_{|M|})$, where $\forall i, 1 \le i \le |M|$, $T_i \subseteq [n], |T_i| = r$, and $\forall i, j, 1 \le i < j \le |M|$, $T_i \cap T_j = \varnothing$.

Definition 1.5. For any $i \in [k]$, let $\mathbf{e}_i \in \mathbb{F}_q^k$ be the i-th standard basis vector.

Theorem 1.1 (Structure Theorem)**.** If $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is an (r, δ, ε)-LDC such that $\forall i \in [n], E_i(\mathbf{x}) = \langle \mathbf{v}_i, \mathbf{x} \rangle$, that is, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{F}_q^k$ are the rows of the generating matrix of E, then $\forall i \in [k]$, $\exists$ an r-matching $M^i = (T_1^i, T_2^i, \ldots, T_{m_i}^i), m_i \geq \frac{\delta n}{r}$, such that $\forall j \in [m_i]$, $\mathbf{e}_i \in \text{span}\{\mathbf{v}_l | l \in T_j^i\}$. That is, each tuple in M^i spans $\mathbf{e}_i$ so it can be used to recover $\mathbf{x}_i = \langle \mathbf{e}_i, \mathbf{x} \rangle$.

Proof. For each i, there is a distribution μ_i on the r-tuples that are subsets of $[n]$ defined by $D(\mathbf{y}, i)$. Since E is a Locally Decodable Code, $\forall \mathbf{x} \in \mathbb{F}_q^k, \forall \mathbf{y} \in \mathbb{F}_q^n$ with $dist(\mathbf{y}, E(\mathbf{x})) \leq \delta n$, $Pr[D(\mathbf{y}, i) = \mathbf{x}_i] \geq 1 - \varepsilon$. Then for a fixed i, we have the following (we start out with no error in the encoded message and we will add the error gradually):

$$\mathbb{E}_{T \sim \mu_i}[D(E(\mathbf{x}), i) = \mathbf{x}_i | D \text{ reads } T] \geq 1 - \varepsilon.$$

Now we randomize uniformly over $\mathbf{x} \sim \mathbb{F}_q^k$, and we get:

$$\mathbb{E}_{T \sim \mu_i} Pr_{\mathbf{x} \sim \mathbb{F}_q^k}[D(E(\mathbf{x}), i) = \mathbf{x}_i | D \text{ reads } T] \geq 1 - \varepsilon.$$

Let us define $P_T = Pr_{\mathbf{x} \sim \mathbb{F}_q^k}[D(E(\mathbf{x}), i) = \mathbf{x}_i | D \text{ reads } T]$ for any $T \sim \mu_i$.

Then there exists $T \subseteq [n]$ such that $P_T \geq 1 - \varepsilon > \frac{1}{q}$. This means that there is a probabilistic function f that takes $\{\langle \mathbf{v}_j, \mathbf{x} \rangle, j \in T\}$ as input and returns $\langle \mathbf{e}_i, \mathbf{x} \rangle$ with probability more than $\frac{1}{q}$ (the function is computed by D when it reads T). But this can only happen if $\mathbf{e}_i \in \text{span}\{\mathbf{v}_j, j \in T\}$. To see why, consider some fixed $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k \in \mathbb{F}_q^k$ such that $\mathbf{w}_k \notin \text{span}\{w_i | 1 \leq i \leq k-1\}$, and some $\mathbf{x} \in \mathbb{F}_q^k$ chosen uniformly at random so that $\langle \mathbf{w}_1, \mathbf{x} \rangle, \langle \mathbf{w}_2, \mathbf{x} \rangle, \ldots, \langle \mathbf{w}_{k-1}, \mathbf{x} \rangle$ are fixed, then $\langle \mathbf{w}_k, \mathbf{x} \rangle$ is still uniformly distributed, therefore it cannot be predicted with probability better than $\frac{1}{q}$.

We have found one of the sets in the matching. Let $T_1 = T$ and continue as follows. Let $E(\mathbf{x})|_{\sim T}$ denote the random variable obtained from $E(\mathbf{x})$ by assigning uniformly random values to entries in T. As before, since $|T| \leq \delta n$, we have:

$$\mathbb{E}_{T \sim \mu_i} Pr_{\mathbf{x} \in \mathbb{F}_q^k}[D(E(\mathbf{x})|_{\sim T_1}, i) = \mathbf{x}_i | D \text{ reads } T] \geq 1 - \varepsilon.$$

Set $P'_T = Pr_{\mathbf{x} \in \mathbb{F}_q^k}[D(E(\mathbf{x})|_{\sim T_1}, i) = \mathbf{x}_i | D \text{ reads } T]$. From the inequality above it follows that there is a $T_2 \sim \mu_i$ such that $P'_{T_2} \geq 1 - \varepsilon > \frac{1}{q}$. Without loss of generality, $T_2 \cap T_1 = \varnothing$, since entries in T_1 are random noise and do not help us in finding $\mathbf{x}_i$ (so we can just ignore them). In this manner, some sets might end up having size smaller than r, in which case we can complete them by adding arbitrary coordinates to them. Using the same argument as above, $\mathbf{e}_i \in \text{span}\{\mathbf{v}_j | j \in T_2\}$.

This process can continue until $|\bigcup_{j=1}^m T_j| > \delta n$, because before that the encoded message has no more than δn errors. Thus $m_i \geq \frac{\delta n}{r}$.

□

Theorem 1.2 (Converse of the Structure Theorem)**.** Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}_q^k$ be such that $\forall i \in [k], \exists$ an r-matching M^i of size $m_i = \alpha n$ so that $\forall T \subseteq M^i$, $\mathbf{e}_i \in \text{span}\{\mathbf{v}_j | j \in T\}$. Then $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ defined by $E_j(\mathbf{x}) = \langle \mathbf{x}, \mathbf{v}_j \rangle$ is an (r, δ, ε)-LDC $\forall \delta, \varepsilon$ such that $\delta \leq \varepsilon \alpha$.

Proof. The decoder $D(\mathbf{y}, i)$ for E works in the following way. It first picks an r-tuple T uniformly at random from M^i, then queries $\{\mathbf{y}_j | j \in T\}$. Now since $\mathbf{e}_i \in \text{span}\{\mathbf{v}_j | j \in T\}$, there exist coefficients $\{a_j | j \in T\}$ such that $\mathbf{e}_i = \sum_{j \in T} a_j \mathbf{v}_j$. Then notice that

$$\begin{aligned}\mathbf{x}_i &= \langle \mathbf{x}, \mathbf{e}_i \rangle \\ &= \langle \mathbf{x}, \sum_{j \in T} a_j \mathbf{v}_j \rangle \\ &= \sum_{j \in T} \langle \mathbf{x}, a_j \mathbf{v}_j \rangle \\ &= \sum_{j \in T} a_j \langle \mathbf{x}, \mathbf{v}_j \rangle \\ &= \sum_{j \in T} a_j E_j(\mathbf{x}).\end{aligned}$$

So after querying $\{\mathbf{y}_j | j \in T\}$, D returns $\sum_{j \in T} a_j \mathbf{y}_j$. A potential error can only come from an error in $\mathbf{y}$. There are at most δn errors in $\mathbf{y}$ by definition, which means that at most δn out of m_i members of M^i can contain an error. Thus, if we're picking an r-tuple uniformly at random from M^i, then the probability of hitting a member that contains an error is at most $\frac{\delta n}{m_i} = \frac{\delta}{\alpha} \leq \varepsilon$.

□

This means that up to a factor of $\frac{1}{r}$ for δ, the two definitions of LDCs are equivalent. Since we will mostly be interested in $r = O(1)$, this is a minor loss.

Definition 1.6. We will say that a code $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is an r-LDC given in *matching form* if it is specified by k r-matchings $M^1, \ldots, M^k$ such that $\forall i, T \in M^i$, $\mathbf{e}_i \in \text{span}\{\mathbf{v}_j | j \in T\}$, where $(\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \left(\mathbb{F}_q^k\right)^n$ are the rows of the generating matrix of E. That is, an r-LDC can simply be given as a set of vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \left(\mathbb{F}_q^k\right)^n$ with k r-matchings $M^1, \ldots, M^k$ (δ and ε are given by $\delta = \frac{\varepsilon}{\alpha}$).

Definition 1.7. We will say that a code $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ with rows of the generating matrix $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a code *without repetitions*, if $\forall i \neq j, \mathbf{v}_i \neq \mathbf{v}_j$. We will say that a code $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ with rows of the generating matrix $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a code *with repetitions* if it is possible for it to not be without repetitions.

Exercise 1.4. Show that if $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is a $(1, \delta, \varepsilon)$-LDC then $\delta \leq \frac{1}{k}$. Show a code matching these parameters.

Exercise 1.5. In the Hadamard code construction, the coefficients used in the linear combinations used for decoding are 1 and -1. Show that from any $(2, \delta, \varepsilon)$-LDC E_1 we can construct a $(2, \delta', \varepsilon')$-LDC E_2 that has this property, where $\delta' = O(\delta)$ and $\varepsilon' = O(\varepsilon)$.

Hint. $\forall \mathbf{v} \in \mathbb{F}_q^k$, let $\tilde{\mathbf{v}} = \lambda \mathbf{v}$ such that for $i_{min} = \min_{i \in [k], \mathbf{v}_i \neq 0}\{i\}$, $\tilde{\mathbf{v}}_{i_{min}} = 1$, replace $\mathbf{v}_1, \ldots, \mathbf{v}_n$ from E_1 with $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_n$ in E_2.

Exercise 1.6. Suppose $E : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is a code given by $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}_q^k$ and r-matchings $M^1, \ldots, M^k$, so that $\frac{1}{k} \sum_{i=1}^k |M^i| \geq \frac{\delta n}{r}$ for some δ (instead of each $|M^i| \geq \frac{\delta n}{r}$). Show how to use E to construct an $(r, \delta', \varepsilon')$-LDC $E' : \mathbb{F}_q^{\frac{k}{2}} \to \mathbb{F}_q^n$, where $\delta' = O(\delta)$ and $\varepsilon' = O(\delta)$.

Exercise 1.7. Consider the Hadamard code over $E : \mathbb{F}_2^k \to \mathbb{F}_2^{2^k}$. Suppose that the goal of the decoder is to recover the sum of message bits $\sum_{i=1}^k \mathbf{x}_i \in \mathbb{F}_2$ locally, by making at most 2 queries. Show how this can be done. What about other linear combinations? Will this still work in $\mathbb{F}_q, q > 2$?