Lecture 13: LDC lower bounds

Sivakanth Gopi

November 6, 2019

In the last few lectures, we have seen constructions of LDCs (and LCCs). In the next few lectures, we will look at lower bounds on the length of LDCs. Let $C : \{0,1\}^k \to \{0,1\}^n$ be a (q, δ, η)-LDC. We want to prove lower bounds on the length n as a function of k, q, δ, η. We are mainly interesting in the regime where q, δ, η are some fixed constants and k is growing. Firstly, it is not hard to see that 1-query LDCs (over constant size alphabet) do not exist over large lengths. Before we prove lower bounds for $q \geq 2$, we will prove some structural results for LDCs.

Smooth codes

Katz and Trevisan [KT00] observed that LDC decoders must have the property that they select their queries according to distributions that do not favor any particular coordinate. The intuition for this is that if they did favor a certain coordinate, then corrupting that coordinate would cause the decoder to err with too high a probability. If instead, queries are sampled according to a "smooth" distribution, they will all fall on uncorrupted coordinates with good probability provided the fraction of corrupted coordinates δ and query complexity q aren't too large. Note that, we can always assume that the marginal distribution of each query is identical. This is because the decoder can always uniformly permute the queries before making them. The following definitions allows us to make this intuition precise.

Definition 1 (Smooth distribution)**.** *A distribution $\mathcal{D}$ over $[n]$ is called c-smooth if for every $i \in [n]$, $\Pr_{\mathcal{D}}[i] \leq \frac{c}{n}$.*

Definition 2 (Smooth LDC)**.** *Let Σ be some finite alphabet. For positive integers k, n, q and parameters $\eta, c > 0$, a map $C : \{0,1\}^k \to \Sigma^n$ is a (q, c, η)-smooth code if, for every $i \in [k]$, there exists a randomized decoder $\mathcal{A}_i$ such that*

1. *For every $x \in \{0,1\}^k$,*

$$\Pr\left[\mathcal{A}_i\big(C(x)\big) = x_i\right] \geq \frac{1}{2} + \eta. \tag{1}$$

2. *The decoder $\mathcal{A}_i$ (non-adaptively) queries at most q coordinates of $C(x)$.*

3. *The distribution of each query that $\mathcal{A}_i$ makes is c-smooth (as defined in Definition 1).*

When the parameter η is not explicity mentioned, usually it is assumed to be some fixed absolute constant. A $(q, 1, \eta)$-smooth LDC is called a perfectly smooth LDC. In a perfectly smooth LDC, the marginal distribution of each query that the decoder makes is uniform over all the coordinates. The following lemma from [KT00] shows that LDCs and smooth LDCs are closely related.

Note that all the constructions of LDCs we have seen so far are perfectly smooth.

Proposition 3 ([KT00]). *If $C : \{0,1\}^k \to \Sigma^n$ is a (q, δ, η)-LDC, then C is also a $(q, 1/\delta, \eta)$-smooth LDC. Conversely, if $C : \{0,1\}^k \to \Sigma^n$ is a (q, c, η)-smooth code, then C is also a $(q, \delta, \eta - qc\delta)$-LDC.*

Proof. Suppose C is a (q, δ, η)-LDC. Let $c = \frac{1}{\delta}$. Fix some $i \in [k]$.Let $\mathcal{A}_i$ be a decoder for x_i. Let $\mu_1, \ldots, \mu_q$ be distributions of the q queries that $\mathcal{A}_i$ generates. We will construct a c-smooth decoder $\mathcal{D}_i$ from $\mathcal{A}_i$ as follows. Without loss of generality, we can assume that the marginal distributions of $j_1, \ldots, j_q$ are identical by randomly permuting the queries of $\mathcal{A}_i$. We say that $j \in [n]$ is "bad" if $\Pr_\mu[j] > \frac{c}{n}$. It is clear that the number of bad coordinates is at most $(1/c)n = \delta n$. We will first show the probability that a bad coordinate is queried by $\mathcal{A}_i$ is small. Let $z \in \Sigma^n$ be s.t. z and $C(x)$ agree on good coordinates, but $z_j = \sigma$ for every bad coordinate $j \in [n]$ where $\sigma \in \Sigma$ is some fixed arbitrary symbol. Clearly $\Delta(z, C(x)) \leq \delta n$. Therefore $\Pr[\mathcal{A}_i(z) = x_i] \geq \frac{1}{2} + \eta$.

$\mathcal{D}_i$ simulates $\mathcal{A}_i$ to generate q queries $(j_1, \ldots, j_q) \in [n]$. But $\mathcal{D}_i$ only queries the good coordinates among $j_1, \ldots, j_q$. For every $\ell \in [q]$ s.t. j_ℓ is a bad coordinate, $\mathcal{D}_i$ will not query j_ℓ, but instead assumes that the symbol at j is σ. Otherwise, $\mathcal{D}_i$ makes all the queries and uses $\mathcal{A}_i$ to decode x_i. Therefore $\mathcal{D}_i(C(x))$ has exactly the same distribution as $\mathcal{A}_i(z)$. Therefore $\Pr[\mathcal{D}_i(C(x)) = x_i] \geq \frac{1}{2} + \eta$. Morever by construction, the decoder $\mathcal{D}_i$ is c-smooth.

We will now prove the converse. Suppose $\mathcal{D}_i$ is a c-smooth decoder and let $y \in \Sigma^n$ be such that $\Delta(z, C(x)) \leq \delta n$. Then

$$\Pr[\mathcal{D}_i(z) = x_i] \geq \Pr[\mathcal{D}_i(C(x)) = x_i] - \Pr[\mathcal{D}_i \text{ queries some } j \in [n] \text{ s.t. } z_j \neq C(x)_j].$$

Since each query of $\mathcal{D}_i$ follows a c-smooth distribution, the probability $\mathcal{D}_i$ queries a corrupted coordinate is at most $q \cdot (c/n) \cdot (\delta n) \leq qc\delta$. This proves the converse.

□

We will prove that a smooth LDC needs to have, for each $i \in [k]$, a large matching of q-tuples M_i from which we can decode x_i.

A q-matching is a q-uniform hypergraph with vertex disjoint hyperedges (which are q-tuples). When q is clear from context, we will just call them matchings. The size of a matching is the number of hyperedges.

Lemma 4. *Let $C : \{0,1\}^k \to \Sigma^n$ be (q, c, η)-smooth LDC. For every $i \in [k]$, there exists a q-matching M_i of size $|M_i| \geq (\eta/cq)n$ s.t. for every q-tuple $S \in M_i$,*

$$\Pr_{x \in \{0,1\}^k}[x_i = \mathcal{D}_i(C(x)) \mid \mathcal{D}_i \text{ queries } S] \geq \frac{1}{2} + \frac{\eta}{2}.$$

Note that the probability is over a random message $x \in \{0,1\}^k$.

Proof. Say that a q-tuple S is good (for $\mathcal{D}_i$) if

$$\Pr_{x\in\{0,1\}^k}[x_i = \mathcal{D}_i(C(x)) \mid \mathcal{D}_i \text{ queries } S] \geq \frac{1}{2} + \frac{\eta}{2}.$$

Let H_i be the hypergraph of all the good edges. We will show that H_i contains a large matching M_i of required size.

We will first show that $\mathcal{D}_i$ will query an edge in H_i with probability at least η.

$$\begin{aligned}
\frac{1}{2} + \eta &\leq \Pr\left[\mathcal{D}_i(C(x)) = x_i\right] \\
&\leq \Pr\left[\mathcal{D}_i(C(x)) = x_i \mid \mathcal{D}_i \text{ queries from } H_i\right]\Pr\left[\mathcal{D}_i \text{ queries from } H_i\right] \\
&\quad + \Pr\left[\mathcal{D}_i(C(x)) = x_i \mid \mathcal{D}_i \text{ doesn't query from } H_i\right](1 - \Pr[\mathcal{D}_i \text{ queries from } H_i]) \\
&\leq \Pr\left[\mathcal{D}_i \text{ queries from } H_i\right] + (1/2 + \eta/2)\,(1 - \Pr[\mathcal{D}_i \text{ queries from } H_i])
\end{aligned}$$

This implies that $\Pr[\mathcal{D}_i \text{ queries from } H_i] \geq \eta$. Let M_i be maximal matching in H_i. The vertices in M_i will form a vertex cover for H_i of size $q|M_i|$. Because of smoothness a $\mathcal{D}_i$, the probability of querying a coordinate in this vertex cover is atmost $(c/n)q|M_i|$. Therefore $(c/n)q|M_i| \geq \eta$, which implies that $|M_i| \geq \eta n/(cq)$. □

Katz-Trevisan lower bound: Random restrictions

The first bound we prove is due to Katz and Trevisan [KT00] who also introduced LDCs in the same paper. The idea is that a small random subset of codeword coordinates (n^ε for some $\varepsilon < 1$) should contain information about most ($\Omega(k)$) of the message bits. By information theoretic arguments, we can argue that this implies that $n \gtrsim k^{1/eps}$.

Before that we need the following lemma.

Lemma 5. *Let M be some fixed q-matching of size $|M| \geq \delta n$ over n vertices. If S is a random subset of $[n]$ where each element is chosen independently with probability $p = (4\delta n)^{-1/q}$, then the probability S contains a edge of M is at least $1/4$.*

Proof. Let $t = |M| \geq \delta n$ and let $e_1, \ldots, e_t$ be the edges of M (which will be vertex disjoint). Let Z_i be the indicator random variable that $e_i \in S$ and let $Z = \sum_{i=1}^{t} Z_i$. Then $\Pr[S \text{ hits an edge of } M] = \Pr[Z \neq 0.]$. By Chebychev inequality

$$\Pr[Z = 0] \leq \frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2} \leq \frac{tp^q + \binom{t}{2}p^{2q}}{(tp^q)^2} \leq 3/4.$$

□

Theorem 6 ([KT00]). *A (q, c, η)-smooth LDC $C : \{0,1\}^k \to \Sigma^n$ must have $n \geq_{q,c,\eta} k^{1+1/(q-1)} \log |\Sigma|$.*

Proof. Let S be a random subset of $[n]$ where each element of chosen independently with probability p. Let X be uniformly distributed over $\{0,1\}^k$ and $Y = C(x)|_S$ be the restriction of Y to S. Let $M_1, \ldots, M_k$ be the matchings given by Lemma 4. By Lemma 5, for each M_i, S will contain an edge of M_i with probability at least $1/4$. Therefore $I(X_i, Y) \gtrsim 1$. Therefore $\sum_{i=1}^{k} I(X_i, Y) \gtrsim k$.

We now claim that $I(X;Y) \geq \sum_{i=1}^{k} I(X_i;Y)$. By chain rule of mutual information, $I(X;Y) = \sum_{i=1}^{k} I(X_i;Y|X_{<i})$. Since $X_i, X_{<i}$ are independent, we have

$$I(X_i;Y) \leq I(X_i;Y, X_{<i}) = I(X_i : X_{<i}) + I(X_i;Y|X_{<i}) = I(X_i;Y|X_{<i}).$$

Therefore $I(X;Y) \geq \sum_{i=1}^{k} I(X_i;Y)$.

We are now done since, $I(X;Y) \leq H(Y) = H(C(x)_S) = \sum_{i=0}^{n} \Pr[|S| = i] H(C(x)|_S \mid |S| = i) \leq \sum_{i=1}^{n} \Pr[|S| = i] i \log |\Sigma| = \mathbb{E}[|S|] \log |\Sigma| = pn \log |\Sigma|$. Combining the upper and lower bounds on $I(X;Y)$, we get $k \lesssim pn \log |\Sigma| \lesssim_{q,c,\eta} k^{1+1/(q-1)} \log |\Sigma|$. □

Theorem 6 implies that 2-query LDCs should have $n \gtrsim k^2$. The best construction of 2-query LDCs we know is the Hadamard code which has length $n = 2^k$. What is the truth? In the next class, we will show that Hadamard codes are actually optimal 2-query LDCs!

References

[KT00] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *Proceedings of the 32nd annual ACM symposium on Theory of computing (STOC 2000)*, pages 80–86. ACM Press, 2000.