Lecture 14: Exponential 2-query LDC lower bound

Sivakanth Gopi

November 13, 2019

In the last lecture, we proved the Katz-Trevisan lower bound which shows that a (q, δ, η)-LDC $C : \{0, 1\}^k \to \Sigma^n$ should have

$$n \gtrsim_{q,\delta,\eta} (k/\log |\Sigma|)^{1+1/(q-1)}.$$

We did this by showing that there is a small subset of coordinates which have information about most of the message coordinates. In this lecture, we will assume that $\Sigma = \{-1, 1\}$ unless otherwise stated. For 2-query LDCs, this shows that $n \gtrsim_{q,\delta,\eta,\Sigma} k^2$. Whereas, the best construction of 2-query LDCs (which are also 2-query LCCs) we have seen is the Hadamard code which has $n = 2^k$. We will prove in this lecture, that Hadamard code is nearly optimal!

Linear 2-query LDCs

We will first show an exponential lower bound for linear 2-query LDCs. We will need the following edge isoperimetric inequality for the hypercube.

Lemma 1 (Edge isopermetric inequality for the hypercube). *Let $G = (\mathbb{F}_2^n, E)$ be the hypercube graph where $(x, y) \in E$ iff x, y differ in exaclty one coordinate. Then for every subset $S \subset \mathbb{F}_2^k$,*

$$E(S, S) \le \frac{1}{2}|S| \log_2 |S|,$$

where $E(S, S)$ is the number of edges in G with both endpoints in S.

The sets S for which the above inequality is tight are precisely the subcubes i.e. sets of the form $S = \{x \in \mathbb{F}_2^k : x|_A = a\}$ for some subset $A \subset [n]$.

Theorem 2 ([GKST06]). *If $C : \mathbb{F}_2^k \to \mathbb{F}_2^n$ is a linear 2-query LDC which can tolerate δn corruptions, then $n \ge \exp(\Omega(\delta n))$.*

Proof. Let $C(x) = (\langle v_j, x\rangle)_{j\in[n]}$ for some $v_1, v_2, \ldots, v_n \in \mathbb{F}_2^k$. We have proved in the last class that, for every $i \in [k]$, there exists a matching M_i on $[n]$ of size $|M_i| \ge \delta n$ such that for every edge $(j, j') \in M_i$, we can decode x_i from $C(x)_j$ and $C(x)_{j'}$. Since C is linear, this implies that $x_i = C(x)_j + C(x)_{j'}$. □

Non-linear 2-query LDCs

Theorem 3 ([KW04]). *Let $C : \{-1,1\}^k \to \{-1,1\}^n$ be a $(2,\delta,\eta)$-LDC, then*

$$n \geq \exp\left(\Omega(\delta\eta^2 k)\right).$$

The original proof of Kerenedis and de Wolf used quantum information theory to prove Theorem 3. We will prove it in a different way using matrix concentration bounds. It turns out that there are deep connections between quantum information theory and matrix concentration bounds. So it is possible that the two proofs which look different superficially are indeed the same!

Matrix concentration bounds

Let $a_1, \dots, a_k \in \mathbb{R}$ and let $x \in \{-1,1\}^k$ be uniformly random, then

$$\mathbb{E}_x\left[\left|\sum_{i=1}^k x_i a_i\right|\right] \leq \sqrt{\sum_{i=1}^k a_i^2}.$$

We want an analogous inequality for matrices. Let A be an $n \times n$ matrix over the reals. The spectral norm of A denoted by $\|A\|_{S_\infty}$ is defined as:

$$\|A\|_{S_\infty} = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_{\ell_2}} = \sup_{x,y \neq 0} \frac{y^T A x}{\|y\|_{\ell_2}\|x\|_{\ell_2}}.$$

The spectral norm is also the largest singular value of the matrix A. The following proposition is the analogue of this fact for matrices first proved in [TJ74] (where it was called non-commutative Khintchine inequality).

Proposition 4 (Tomczak-Jaegermann). *Let $A_1, \cdots, A_k$ be $n \times n$ matrices over the reals, then*

$$\mathbb{E}_{x \in \{-1,1\}^k}\left[\left\|\sum_{i=1}^k x_i A_i\right\|_{S_\infty}\right] \lesssim \sqrt{\log n}\left(\sum_{i=1}^k \|A_i\|_{S_\infty}^2\right)^{1/2}$$

where the expectation is over a uniformly random $x \in \{-1,1\}^k$.

See [Tro15, Theorem 4.1.1] for the statement above and [Tro15] for more on such matrix concentration inequalities. From 4, and general concentration tools like the Lipchitz concetration theorem, one can show exponential tail bounds for $\left\|\sum_{i=1}^k x_i A_i\right\|_{S_\infty}$.

Proof of lower bound

We will not worry about the dependence on η, δ. But by being more careful, one can get the exact bound in Theorem 3 as shown in [Gop18]. In the last lecture, we have shown that we can assume that, on an average codeword, the decoders sample a random edge of a large matching and query the vertices of that edge.

Lemma 5. *Let $C : \{-1,1\}^k \to \{-1,1\}^n$ be $(2, \delta, \eta)$-LDC. For every $i \in [k]$, there exists a matching M_i of size $|M_i| \gtrsim_{\eta,\delta} n$ s.t. for every edge $(j,k) \in M_i$,*

$$\mathbb{E}_{x\in\{-1,1\}^k}\mathbb{E}_{(j,k)\in M_i}\left[x_i\mathcal{D}_i(C(x)_j, C(x)_k)\right] \geq \eta.$$

We can write any function $f : \{-1,1\}^2 \to \{-1,1\}$ as

$$f(z_1, z_2) = \hat{f}(\phi) + \hat{f}(\{1\})z_1 + \hat{f}(\{2\})z_2 + \hat{f}(\{1,2\})z_1z_2.$$

Such a representation is called the Fourier expansion and the coefficients are called the Fourier coefficients. The Fourier coefficients satisfy $\hat{f}(S) = \mathbb{E}_z[f(z)\prod_{\ell\in S} z_\ell]$ and therefore $|\hat{f}(S)| \leq 1$. Applying this to $f = \mathcal{D}_i$, we can assume WLOG[1] that

$$\left|\mathbb{E}_{x\in\{-1,1\}^k}\mathbb{E}_{(j,k)\in M_i}\left[x_iC(x)_jC(x)_k\right]\right| \gtrsim \eta.$$

In other words, we can assume that the decoders just output the parity of the two bits they have queried or the negation of that. For simplicity, let us assume that they always output the parity of the two bits they queried (in $\{-1,1\}$ notation, they output the product). Therefore,

$$\mathbb{E}_{x\in\{-1,1\}^k}\mathbb{E}_{(j,k)\in M_i}\left[x_iC(x)_jC(x)_k\right] \gtrsim \eta.$$

Let $A_1, A_2, \ldots, A_k$ represent the adjacency matrices of the matchings $M_1, M_2, \ldots, M_k$ respectively. Then we have

$$C(x)^TA_iC(x) = \sum_{(j,k)\in M_i} C(x)_jC(x)_k.$$

Therefore we have $\mathbb{E}_{x\in\{-1,1\}^k}C(x)^TA_iC(x) \gtrsim \eta|M_i| \geq \eta\delta n$. Adding the above inequality for each $i \in [k]$, we get

$$\mathbb{E}_x\left[C(x)^T\left(\sum_{i=1}^k x_iA_i\right)C(x)\right] \gtrsim \eta\delta kn.$$

We can upper bound the LHS of the above inequality as

$$E_x\left[C(x)^T\left(\sum_{i=1}^k x_iA_i\right)C(x)\right] \leq \mathbb{E}_x\left[\left\|\sum_{i=1}^k x_iA_i\right\|_{S_\infty} \cdot \|C(x)\|_{\ell_2}^2\right]$$
$$= n \cdot \mathbb{E}_x\left[\left\|\sum_{i=1}^n x_iA_i\right\|_{S_\infty}\right].$$

[1] It can also be true that x_i is correlated with just $C(x)_j$ or just with $C(x)_k$. This would mean that we can decode x_i with just one query. But there are cannot be many such coordinates, so we can ignore them.

Since the matrix A_i is equivalent to a diagonal matrix with $\{0,1\}$ entries after permuting rows and columns, it is easy to see that $\|A_i\|_{S_\infty} \leq 1$. So Lemma 4 implies that

$$\mathbb{E}_x \left[\left\| \sum_{i=1}^{n} x_i A_i \right\|_{S_\infty} \right] \lesssim \sqrt{\log(n)}\sqrt{k} = \sqrt{k \log n}.$$

Combining the above inequalities, we get

$$n\sqrt{k \log n} \gtrsim \eta\delta kn \Rightarrow n \geq \exp(\Omega(\eta^2\delta^2 k)).$$

2-query LDCs over large alphabet

We want to understand the optimal length of 2-query LDCs over growing alphabet size. 2-query LDCs over large alphabet are intimately related to 2-server private information retrieval schemes (we will see this later in the course). Specifically, we are interested in the regime where $|\Sigma| = n$ i.e. the alphabet size is comparable to the length of the codewords. We have already seen the Katz-Trevisan bound which says that $n \gtrsim (k/\log|\Sigma|)^2$. The exponential lower bound, we have proved in this class can be extended to the growing alphabet case, but it quickly degrades with the size of the alphabet.

A 2-server private information retrieval (PIR) scheme allows a user to retrieve information from two (non-colluding) servers without revealing any information about their query to either server. Private information Retrieval schemes were defined in [CGKS98], before LDCs were even formally defined. In the paper where they defined LDCs [KT00], Katz and Trevisan showed that PIR schemes are closely related to LDCs.

Theorem 6 ([KW04]). *Let $C : \{-1,1\}^k \to \Sigma^n$ be a $(2,\delta,\eta)$-LDC, then*

$$n \geq \exp\left(\Omega\left(\frac{\delta\eta^2 k}{|\Sigma|^2}\right)\right).$$

Open Problem 7. *Suppose we have a 2-query LDC $C : \{-1,1\}^k \to \Sigma^n$ where $|\Sigma| = n$. Is it true that $n = k^{\omega(1)}$?*

The best known construction of a 2-query LDC in this regime (i.e., $|\Sigma| = n$) achieves $n = \exp(k^{o(1)})$ [DG16]. Therefore there can subexponential length LDCs in the large alphabet regime!

Lower bounds for q-query LDCs

We have proved in the last class, the Katz-Trevisan lower bound, which shows that if $C : \{-1,1\}^k \to \{-1,1\}^n$ is a q-query LDC, then $n \gtrsim k^{1+1/(q-1)}$. This has been improved by Kerenidis and de Wolf by a reduction to the exponential 2-query lower bound.

Theorem 8 ([KW04]). *Let $C : \{-1,1\}^k \to \{-1,1\}^n$ be a (q,δ,η)-LDC, then*

$$n \gtrsim_{\delta,\eta} \left(\frac{k}{\log k}\right)^{1+\frac{1}{\lceil q/2 \rceil - 1}}.$$

Here we will sketch the main idea in the reduction. For simplicity, let us prove that a 4-query LDC should have $n \gtrsim (k/\log k)^2$. The main idea is a reduction. Given a 4-query LDC $C : \{-1,1\}^k \to \{-1,1\}^n$, we will construct a 2-query LDC $C' : \{-1,1\}^k \to \{-1,1\}^N$ where $N = n^{\sqrt{n}}$. Now applying the 2-query exponential lower bound, we get the required lower bound. The new code is defined as

$$C'(x) = C(x)^{\otimes \sqrt{n}}.$$

Claim 9. *C' is a 2-query LDC.*

Proof. Homework! □

An approach to improve lower bounds for q-query LDCs

Here is one way to generalize the matrix concentration approach to prove lower bounds for q-query LDCs for $q \geq 3$. Given a q-multilinear form Λ, we define its norm as:

$$\|\Lambda\| = \sup\left\{\Lambda(z_1, \cdots, z_q) : z_1, z_2, \ldots, z_q \in \{-1,1\}^n\right\}. \qquad (1)$$

Let $\mathcal{C} : \{-1,1\}^k \to \{-1,1\}^n$ be a q-query LDC and let $M_1, M_2, \ldots, M_k$ be the decoding q-matchings. For each matching M_i, we can define a q-multilinear form Λ_i as:

$$\Lambda_i(z_1, \cdots, z_q) = \frac{1}{n} \sum_{(j_1, \cdots, j_q) \in \mathcal{M}_i} \prod_{i=1}^{q} (z_i)_{j_i}.$$

So for every $x \in \{-1,1\}^k$,

$$\mathbb{E}_{x \in \{-1,1\}^k} \left[x_i \Lambda_i(C(x), \cdots, C(x))\right] \gtrsim_{\varepsilon,\delta,q} 1.$$

Summing over $i \in [k]$, we have

$$\mathbb{E}_x \left[\left(\sum_{i=1}^{k} x_i \Lambda_i\right)(C(x), \cdots, C(x))\right] \gtrsim_{\varepsilon,\delta,q} k.$$

We can upper bound the LHS as:

$$\mathbb{E}_x \left[\left(\sum_{i=1}^{k} x_i \Lambda_i\right)(C(x), \cdots, C(x))\right] \leq \mathbb{E}_x \left[\left\|\sum_{i=1}^{k} x_i \Lambda_i\right\|\right].$$

Therefore we have

$$\mathbb{E}_x \left[\left\|\sum_{i=1}^{k} x_i \Lambda_i\right\|\right] \gtrsim_{\varepsilon,\delta,q} k.$$

So if we have a statement analogous to Proposition 4, which gives a good upper bound on $\mathbb{E}_x \left[\left\|\sum_{i=1}^{k} x_i \Lambda_i\right\|\right]$, we get good q-query LDC lower bounds. It can be shown that $\mathbb{E}_x \left[\left\| \sum_{i=1}^{k} x_i \Lambda_i \right\| \right] \leq f_q(n)\sqrt{k}$ for some function $f_q(n)$. This would prove the upper bound

$$k \leq f_q(n)^2$$

for q-query LDCs $C : \{-1,1\}^k \to \{-1,1\}^n$. It is trivial to show that $f_q(n) \lesssim \sqrt{n}$. Proposition 4 implies that $f_2(n) \lesssim \sqrt{\log n}$. The existence of subexponential 3-query LDCs [Efr09] implies that $f_3(n) \geq (\log n)^{\Omega(\log \log n)}$. Showing that $f_3(n) \leq n^{1/4-\alpha}$ for some $\alpha > 0$ implies a super-quadratic lower bound for 3-query LDCs which is currently not known.

Summary of known results

The following tables show the summary of best known constructions and lower bounds for q-query LDCs/LCCs $C : \{0,1\}^k \to \{0,1\}^n$. δ, η are assumed to be some fixed constants. Smaller order terms in k like $\log k$ and $\log \log k$ are also ignored.

	$q = 2$		$q = O(1),\ q \geq 3$	
	Upper bounds	Lower bounds	Upper bounds	Lower bounds
LDCs	$n \leq 2^k$ Hadamard Code	$n \geq \exp(\Omega(k))$	$n = \exp(k^{o(1)})$ Matching Vector Codes	$n \gtrsim k^{1+\frac{1}{\lceil q/2 \rceil - 1}}$
LCCs	"	"	$n = \exp\left(O_q(k^{1/(q-1)})\right)$ Reed-Muller codes	"

	$q = (\log n)^t,\ t = O(1),\ t > 1$		$q = n^{o(1)}$	
	Upper bounds	Lower bounds	Upper bounds	Lower bounds
LDCs	$n \lesssim k^{1+\frac{1}{t-1}}$ Reed-Muller codes	$n \geq \Omega(k)$ (Trivial)	$n = O(k)$ Match GV bound [KRR+19]	$n = \Omega(k)$ (Trivial)
LCCs	"	"	"	"

References

[CGKS98] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. *Journal of the ACM*, 45(6):965–981, 1998.

[DG16] Zeev Dvir and Sivakanth Gopi. 2-Server PIR with subpolynomial communication. *J. ACM*, 63(4):39:1–39:15, September 2016. Preliminary version appeared in STOC 2015.

[Efr09] Klim Efremenko. 3-query locally decodable codes of subexponential length. In *STOC*, pages 39–44, 2009.

[GKST06] Oded Goldreich, Howard Karloff, Leonard J Schulman, and Luca Trevisan. Lower bounds for linear locally decodable codes and private information retrieval. *Computational Complexity*, 15(3):263–296, 2006.

[Gop18] Sivakanth Gopi. *Locality in Coding Theory*. PhD thesis, Princeton University, 2018.

[KRR$^+$19] Swastik Kopparty, Nicolas Resch, Noga Ron-Zewi, Shubhangi Saraf, and Shashwat Silas. On list recovery of high-rate tensor codes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA*, pages 68:1–68:22, 2019.

[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.

[KW04] Iordanis Kerenidis and Ronald de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. *J. of Computer and System Sciences*, 69:395–420, 2004. Preliminary version appeared in STOC'03.

[TJ74] Nicole Tomczak-Jaegermann. The moduli of smoothness and convexity and the rademacher averages of the trace classes S_p, $1 \leq p \leq \infty$. *Studia Mathematica*, 50(2):163–182, 1974.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.