Electronic Colloquium on Computational Complexity, Report No. 101 (2022)

A Near-Cubic Lower Bound for 3-Query Locally Decodable Codes from Semirandom CSP Refutation

Omar Alrabiah*
oalrabiah@berkeley.edu
UC Berkeley

Venkatesan Guruswami†
venkatg@berkeley.edu
UC Berkeley

Pravesh K. Kothari‡
praveshk@cs.cmu.edu
Carnegie Mellon University

Peter Manohar§
pmanohar@cs.cmu.edu
Carnegie Mellon University

Abstract

A code $C\colon \{0,1\}^k \to \{0,1\}^n$ is a q-locally decodable code (q-LDC) if one can recover any chosen bit b_i of the message $b \in \{0,1\}^k$ with good confidence by randomly querying the encoding $x := C(b)$ on at most q coordinates. Existing constructions of 2-LDCs achieve $n = \exp(O(k))$, and lower bounds show that this is in fact tight. However, when $q = 3$, far less is known: the best constructions achieve $n = \exp(k^{o(1)})$, while the best known results only show a quadratic lower bound $n \geq \widetilde{\Omega}(k^2)$ on the blocklength.

In this paper, we prove a near-cubic lower bound of $n \geq \widetilde{\Omega}(k^3)$ on the blocklength of 3-query LDCs. This improves on the best known prior works by a *polynomial* factor in k. Our proof relies on a new connection between LDCs and refuting constraint satisfaction problems with limited randomness. Our quantitative improvement builds on the new techniques for refuting *semirandom* instances of CSPs developed in [GKM22] and, in particular, relies on bounding the $(\infty \to 1)$-norm of appropriate *Kikuchi* matrices.

*Supported in part by a Saudi Arabian Cultural Mission (SACM) Scholarship, NSF CCF-2228287 and V. Guruswami's Simons Investigator Award.

†Supported in part by NSF grants CCF-CCF-2228287 and CCF-2211972 and a Simons Investigator award.

‡Supported in part by an NSF CAREER Award #2047933, a Google Research Scholar Award, and a Sloan Fellowship.

§Supported in part by an ARCS Scholarship, NSF Graduate Research Fellowship (under grant numbers DGE1745016 and DGE2140739), and NSF CCF-1814603.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ISSN 1433-8092

Contents

1 Introduction

A binary *locally decodable code* (LDC) $C\colon \{0,1\}^k \to \{0,1\}^n$ maps a k-bit message $b \in \{0,1\}^k$ to an n-bit codeword $x \in \{0,1\}^n$ with the property that the receiver, when given oracle access to $y \in \{0,1\}^n$ obtained by corrupting x in a constant fraction of coordinates, can recover any chosen bit b_i of the original message with good confidence by only querying y in a few locations. More formally, a code C is q-locally decodable if for any input $i \in [k]$, the decoding algorithm makes at most q queries to the corrupted codeword y and recovers the bit b_i with probability $1/2 + \varepsilon$, provided that $\Delta(y, C(b)) \coloneqq |\{v \in [n] : y_v \neq C(b)_v\}| \leq \delta n$, where δ, ε are constants. Though formalized later in [KT00], locally decodable codes were instrumental in the proof of the PCP theorem [AS98, ALM+98], and have deep connections to many other areas of complexity theory (see Section 7 in [Yek12]), including worst-case to average-case reductions [Tre04], private information retrieval [Yek10], secure multiparty computation [IK04], derandomization [DS05], matrix rigidity [Dvi10], data structures [Wol09, CGW10], and fault-tolerant computation [Rom06].

A central research focus in coding theory is to understand the largest possible *rate* achievable by a q-query locally decodable code. For the simplest non-trivial setting of $q = 2$ queries, we have a complete understanding: the Hadamard code provides an LDC with a blocklength $n = 2^k$ and an essentially matching lower bound of $n = 2^{\Omega(k)}$ was shown in [KW04, GKST06, Bri16, Gop18].

In contrast, there is a wide gap in our understanding of 3 or higher query LDCs. The best known constructions are based on families of *matching vector codes* [Yek08, Efr09, DGY11] and achieve $n = 2^{k^{o(1)}}$. In particular, the blocklength is slightly subexponential in k and asymptotically improves on the rate achievable by 2-query LDCs. The known lower bounds, on the other hand, are far from this bound. The first LDC lower bounds are due to Katz and Trevisan [KT00], who proved that q-query LDCs require a blocklength of $n \geq \Omega(k^{\frac{q}{q-1}})$. This was later improved in 2004 by Kerenedis and de Wolf [KW04] via a "quantum argument" to obtain $n \geq k^{\frac{q}{q-2}}/\mathrm{polylog}(k)$ when q is even, and $n \geq k^{\frac{q+1}{q-1}}/\mathrm{polylog}(k)$ when q is odd. For the first nontrivial setting of $q = 3$, their result yields a nearly quadratic lower bound of $n \geq \Omega(k^2/\log^2 k)$ on the blocklength. Subsequently, Woodruff [Woo07, Woo12] improved this bound by $\mathrm{polylog}(k)$ factors to obtain a lower bound of $n \geq \Omega(k^2/\log k)$ for non-linear codes, and $n \geq \Omega(k^2)$ for linear codes. Very recently, Bhattacharya, Chandran, and Ghoshal [BCG20] used a combinatorial method to give a new proof of the quadratic lower bound of $n \geq \Omega(k^2/\log k)$, albeit with a few additional assumptions on the code.

Our Work. In this work, we show a near-cubic lower bound $n \geq k^3/\mathrm{polylog}(k)$ on the blocklength of any 3-query LDC. This improves on the previous best lower bound by a $\tilde{O}(k)$ factor. More precisely, we prove:

Theorem 1. *Let $C\colon \{0,1\}^k \to \{0,1\}^n$ be a code that is $(3, \delta, \varepsilon)$-locally decodable. Then, it must hold that $k^3 \leq n \cdot O((\log^{14} n)/\varepsilon^{32}\delta^{16})$. In particular, if δ, ε are constants, then $n \geq k^3/\mathrm{polylog}(k)$.*

We have not attempted to optimize the dependence on $\log n$, ε, and δ in Theorem 1. We also suspect that it is simple to extend Theorem 1 to nonbinary alphabets, with a polynomial loss in the alphabet size. Finally, using known relationships between locally correctable codes (LCCs) and LDCs (e.g., Theorem A.6 of [BGT17]), Theorem 1 implies a similar lower bound for 3-query LCCs.

Our main tool is a new connection between the existence of locally decodable codes and refutation of instances of Boolean CSPs with limited randomness. This connection is similar in spirit to the connection between PCPs and hardness of approximation for CSPs, in which one produces a q-ary CSP from a PCP with a q-query verifier by adding, for each possible query set of the verifier, a local constraint that asserts that the verifier accepts when it queries this particular set. To refute the resulting CSP instance, our proof builds on the spectral analysis of *Kikuchi matrices* employed in the recent work of [GKM22], which obtained strong refutation algorithms for semirandom and smoothed CSPs and proved the hypergraph Moore bound conjectured by Feige [Fei08].

Up to polylog(k) factors, the best known lower bound of $n \geq k^{\frac{q+1}{q-1}}/\text{polylog}(k)$ for q-LDCs for odd q can be obtained by simply observing that a q-LDC is also a $(q+1)$-LDC, and then invoking the lower bound for $(q+1)$-query LDCs. Our improvement for $q = 3$ thus comes from obtaining the same tradeoff with q as in the case of even q, but now for $q = 3$. For technical reasons, our proof does not extend to odd $q \geq 5$; we briefly mention in Section 1.1 the place where the natural generalization fails. We leave proving a lower bound of $n \geq k^{\frac{q}{q-2}}/\text{polylog}(k)$ for all *odd* $q \geq 5$ as an intriguing open problem.

1.1 Proof overview

The key insight in our proof is to observe that for any q, a q-LDC yields a collection of q-XOR instances, one for each possible message. And, a typical instance has a high value, i.e., there's an assignment that satisfies $\frac{1}{2} + \varepsilon$-fraction of the constraints. To prove a lower bound on the blocklength n for 3-LDCs, it is then enough to show that for any purported construction with $n \ll k^3$, the associated 3-XOR instance corresponding to a uniformly random message has a low value. We establish such a claim by producing a refutation (i.e., a certificate of low value), building on tools from the recent work on refuting smoothed instances of Boolean CSPs [GKM22].

For this overview, we will assume that the code C is a *linear* q-LDC. We will also write the code using $\{-1,1\}$ notation, so that $C\colon \{-1,1\}^k \to \{-1,1\}^n$. By standard reductions (Lemma 6.2 in [Yek12]), one can assume that the LDC is in normal form: there exist q-uniform hypergraph matchings $\mathcal{H}_1, \ldots, \mathcal{H}_k$, each with $\Omega(n)$ hyperedges,[1] and the decoding procedure on input $i \in [k]$ simply chooses a uniformly random $C \in \mathcal{H}_i$, and outputs $\prod_{v \in C} x_v$. Because C is linear, when $x = C(b)$ is the encoding of b, the decoding procedure recovers b_i with probability 1. In other words, for any $b \in \{-1,1\}^k$, the assignment $x = C(b)$ satisfies the set of q-XOR constraints $\forall i \in [k], C \in \mathcal{H}_i, \prod_{v \in C} x_v = b_i$.

The XOR Instance. The above connection now suggests the following approach: let $b \in \{-1,1\}^k$ be chosen randomly, and consider the q-XOR instance with constraints $\forall i \in [k], C \in \mathcal{H}_i, \prod_{v \in C} x_v = b_i$. Since C is a linear q-LDC, this set of constraints will be satisfiable for every choice of b. Thus, proving that the instance is unsatisfiable, with high probability for a uniformly random b, implies a contradiction.

One might expect to show unsatisfiability of a q-XOR instance produced by a sufficiently

[1] A q-uniform hypergraph $\mathcal{H}_i$ is a collection of subsets of $[n]$, called hyperedges, each of size exactly q. The hypergraph $\mathcal{H}_i$ is a matching if all the hyperedges are disjoint.

random generation process by using natural probabilistic arguments. Indeed, if the instance was "fully random" (i.e., both $\mathcal{H}_i$'s and b_i's chosen uniformly at random from their domain), or even semirandom (where $\mathcal{H}_i$'s are worst-case but each constraint C has a uniformly random "right hand side" $b_C \in \{-1, 1\}$), then a simple union bound argument suffices to prove unsatisfiability.

The main challenge in our setting is that the q-XOR instances have significantly *limited* randomness even compared to the semirandom setting: all the constraints $C \in \mathcal{H}_i$ share the *same* right hand side b_i. In particular, the q-XOR instance on n variables has $k \ll n$ bits of independent randomness.

We establish the unsatisfiability of such a q-XOR instance above by constructing a subexponential-sized SDP-based certificate of low value. A priori, bounding the SDP value might seem like a rather roundabout route to show unsatisfiability of a q-XOR instance. However, shifting to this stronger target allows us to leverage the techniques introduced in the recent work of [GKM22] on *semirandom* CSP refutation and to show existence of such certificates of unsatisfiability. Despite the significantly smaller amount of randomness in the q-XOR instances produced in our setting, compared to, e.g., semirandom instances, we show that an appropriate adaptation of the techniques from [GKM22] is powerful enough to exploit the combinatorial structure in our instances and succeed in refuting them.

Warmup: the case when q is even. Certifying unsatisfiability of q-XOR instances when q is even is known to be, from a technical standpoint, substantially easier compared to the case when q is odd. As a warmup, we will first sketch a proof of the known lower bound for q-LDCs when q is even, via our CSP refutation approach. A full formal proof is presented in Appendix A.

The refutation certificate is as follows. Let ℓ be a parameter to be chosen later, and let $N := \binom{n}{\ell}$. For a set $C \in \binom{[n]}{q}$,[2] we let $A^{(C)} \in \mathbb{R}^{N \times N}$ be the matrix indexed by sets $S \in \binom{[n]}{\ell}$, where $A^{(C)}(S, T) = 1$ if $S \oplus T = C$, and 0 otherwise, where $S \oplus T$ denotes the symmetric difference of S and T. We note that $S \oplus T = C$ if and only if $S = C_1 \cup Q$ and $T = C_2 \cup Q$, where C_1 is half of the clause C, C_2 is the other half of the clause C, and Q is an arbitrary subset of $[n] \setminus Q$ of size $\ell - q/2$. This matrix $A^{(C)}$ is the Kikuchi matrix (also called symmetric difference matrix) of [WAM19]. We then set $A = \sum_{i=1}^{k} b_i \sum_{C \in \mathcal{H}_i} A^{(C)}$. By looking at the quadratic form $y^\top A y$ where y is defined as $y_S := \prod_{v \in S} x_v$, where $x = C(b)$, it is simple to observe that $\|A\|_2 \geq (\ell/n)^{q/2} \cdot \sum_{i=1}^{k} |\mathcal{H}_i| \geq (\ell/n)^{q/2} \Omega(kn)$.

As each b_i is an independent bit from $\{-1, 1\}$, the matrix A is the sum of k independent, mean 0 random matrices: we can write $A = \sum_{i=1}^{k} b_i A_i$, where $A_i := \sum_{C \in \mathcal{H}_i} A^{(C)}$. We can then bound $\|A\|_2$ using Matrix Bernstein, which implies that $\|A\|_2 \leq O(\Delta)(\ell \log n + \sqrt{k\ell \log n})$, where Δ is the maximum ℓ_1-norm of a row in any A_i. One technical issue is that there are rows with abnormally large ℓ_1-norm, so Δ can be as large as $\Omega(\ell)$; however, using the "row pruning" technique of [GKM22], we can show that only a small fraction of rows/columns have substantially higher-than-average ℓ_1-norm. This allows us to replace Δ with the maximum of 1 and the average ℓ_1-norm of a row, which is $\sim (\ell/n)^{q/2} \cdot |\mathcal{H}_i|$, i.e., $(\ell/n)^{q/2} n$. That is, $\Delta = \max(1, (\ell/n)^{q/2} n)$. Hence, for $\ell \gg n^{1-2/q}$, we can set $\Delta = (\ell/n)^{q/2} n$.

Combining, we thus have that for $\ell \gg n^{1-2/q}$,

$$(\ell/n)^{q/2} \Omega(kn) \leq \|A\|_2 \leq O(\Delta)(\ell \log n + \sqrt{k\ell \log n}) \leq O(1)(\ell/n)^{q/2} n \cdot (\ell \log n + \sqrt{k\ell \log n}) \ ,$$

[2] We use $\binom{[n]}{t}$ to denote the collection of subsets of $[n]$ of size exactly t.

which implies that $k \leq \ell \cdot \text{polylog}(n)$. Taking $\ell = n^{1-2/q}$ to be the smallest possible setting of ℓ for which the above holds, we obtain the desired lower bound.

The case of $q = 3$. When $q = 3$, or more generally when q is odd, the matrices $A^{(C)}$ are no longer well-defined, as the condition $S \oplus T = C$ is never satisfied. A naive attempt to salvage the above approach is to simply allow the columns of $A^{(C)}$ to be indexed by sets of size $\ell + 1$, rather than ℓ. However, this asymmetry in the matrix causes the spectral certificate to obtain a suboptimal dependence in terms of q, leading to a final bound of $k \leq n^{1-2/(q+1)}\text{polylog}(n)$ same as the current state-of-the-art lower bound for odd q. This is precisely the issue that in general makes refuting q-XOR instances for odd q technically more challenging than even q. The asymmetric matrix effectively pretends that q is $q + 1$, and thus obtains the "wrong" dependence on q.

Our idea is to transform a 3-LDC into a 4-XOR instance and then use an appropriate Kikuchi matrix to find a refutation for the resulting 4-XOR instance. The resulting 4-XOR instance is less structured than ones arising out of 4-LDCs: as we shall see, the $\mathcal{H}_i$'s in the 4-XOR instance will no longer be matchings. Nevertheless, our techniques succeed in refuting such instances by exploiting only some elementary combinatorial properties.

Our reduction works as follows. We randomly partition $[k]$ into two sets, L, R, and fix $b_j = 1$ for all $j \in R$. Then, for each *intersecting pair* of constraints C_i, C_j that intersect with $C_i \in \mathcal{H}_i, i \in L$, $C_j \in \mathcal{H}_j, j \in R$, we add the derived constraint $C_i \oplus C_j$ to our new 4-XOR instance, with right hand side b_i.[3] Because the 3-XOR instance was satisfiable, the 4-XOR instance is also satisfiable. Moreover, the 4-XOR instance has $\sim k^2 n$ constraints, as a typical $v \in [n]$ participates in $\sim k$ hyperedges in $\cup_{i=1}^{k} \mathcal{H}_i$, and hence can be "canceled" to form k^2 derived constraints.

The partition (L, R) is a technical trick that allows us to produce $\sim k^2 n$ constraints in the 4-XOR instance while preserving k independent bits of randomness in the right hand sides of the constraints. If we considered *all* derived constraints, rather than just those that cross the partition (L, R), then it would be possible to produce derived constraints where the right hand sides have nontrivial correlations. Specifically, one could produce 3 constraints with right hand sides $b_i b_j, b_j b_t, b_i b_t$, which are pairwise independent but not 3-wise independent. With the partitioning, however, the right hand sides of any two constraints must either be equal or independent, and in particular there are no nontrivial correlations.

The fact that we have produced more constraints in the 4-XOR instance is crucial, as otherwise we could only hope to obtain the same bound as in the $q = 4$ case in the warmup earlier. However, our reduction does not produce an instance with the same structure as a 4-XOR instance arising from a 4-LDC: if we let $\mathcal{H}'_i$ for $i \in L$ denote the set of derived constraints with right hand side b_i, then we clearly can see that $\mathcal{H}'_i$ is not a matching. In fact, the typical size of $\mathcal{H}'_i$ is $\Omega(nk)$, whereas a matching can have at most n/q hyperedges.

Nonetheless, we can still apply the CSP refutation machinery to try to refute this 4-XOR instance. However, because each $\mathcal{H}'_i$ is no longer a matching, the "row pruning step" now only works if we assume that any pair $p = (u, v)$ of vertices appears in at most $\text{polylog}(n)$ hyperedges in the original 3-uniform hypergraph $\cup_{i=1}^{k} \mathcal{H}_i$. But, if we make this assumption, the rest of the proof follows the

[3] If $|C_i \cap C_j| = 2$, then the derived constraint is a 2-XOR constraint, not 4-XOR. This is a minor technical issue that can be circumvented easily, so we will ignore it for the proof overview.

blueprint of the even q case, and we can prove that $n \geq k^3/\text{polylog}(k)$. We note that a recent work [BCG20] managed to reprove that $n \geq k^2/\text{polylog}(k)$ under a similar assumption about pairs of vertices.

Thus, the final step of the proof is to remove the assumption by showing that no pair of vertices can appear in too many hyperedges. Suppose that we do have many "heavy" pairs $p = (u, v)$ that appear in $\gg \log n$ clauses in the original 3-uniform hypergraph $\mathcal{H} := \cup_{i=1}^{k} \mathcal{H}_i$. Now, we transform the 3-XOR instance into a bipartite 2-XOR instance ([AGK21, GKM22]) by replacing each heavy pair p with a new variable y_p. That is, the 3-XOR clause $C = (u, v, w)$ in $\mathcal{H}_i$ now becomes the 2-XOR clause (p, w), where p is a new variable. In other words, the constraint $x_u x_v x_w = b_i$ is replaced by $y_p x_w = b_i$. Each clause in the bipartite 2-XOR instance now uses one variable from the set of heavy pairs, and one from the original set of variables $[n]$. We then show that if there are too many heavy pairs, then this instance has a sufficient number of constraints in order to be refuted, and is thus not satisfiable, which is again a contradiction.

Finally, we note that for larger odd $q \geq 5$, the proof showing that there not too many heavy pairs breaks down, and this is what prevents us from generalizing Theorem 1 to all odd q.

1.2 Discussion: LDCs and the CSP perspective

Prior work on lower bounds for q-LDCs reduce q-query LDCs with even q to 2-query LDCs, and then apply the essentially tight known lower bounds for 2-query LDCs. (To handle the odd q case, they essentially observe that a q-LDC is also a $(q+1)$-LDC.) While the warmup proof we sketched earlier (and present in Appendix A) for even q is in the language of CSP refutation, it is in fact very similar to the reduction from q-LDCs to 2-LDCs for q even used in the proof in [KW04]. The reduction in [KW04] (see also Exercise 4 in [Gop19]) employs a certain tensor product, and while it is not relevant to their argument, the natural matrix corresponding to the 2-LDC produced by their reduction is in fact very closely related to the Kikuchi matrix A of [WAM19].

The main advantage of the CSP refutation viewpoint is that it suggests a natural route to analyze q-LDCs for *odd* q via an appropriately modified Kikuchi matrix. By viewing the 3-LDC as a 3-XOR instance, we obtain a natural way to produce a related 4-XOR instance using a reduction that *does not correspond to a* 4*-LDC*. In fact, if our reduction were to only produce a 4-LDC, then we would not expect to obtain an improved 3-LDC lower bound without improving the 4-LDC lower bound as well. In a sense, this relates to the key strength of the CSP viewpoint in that it arguably the "right" level of abstraction. On one hand, it naturally suggests reductions from 3-LDCs to 4-XOR that are rather unnatural if one were to follow the more well-trodden route of reducing odd query LDCs to even query ones. On the other hand, the ideas from semirandom CSP refutation are resilient enough to apply, with some effort, to even the more general, non-semirandom instances arising in such reductions, and so we can still prove lower bounds. Further exploration of such an approach to obtain stronger lower bounds for LDCs is an interesting research direction.

2 Preliminaries

2.1 Basic notation

We let $[n]$ denote the set $\{1, \ldots, n\}$. For two subsets $S, T \subseteq [n]$, we let $S \oplus T$ denote the symmetric difference of S and T, i.e., $S \oplus T := \{i : (i \in S \wedge i \notin T) \vee (i \notin S \wedge i \in T)\}$. For a natural number $t \in \mathbb{N}$, we let $\binom{[n]}{t}$ be the collection of subsets of $[n]$ of size exactly t.

For a rectangular matrix $A \in \mathbb{R}^{m \times n}$, we let $\|A\|_2 := \max_{x \in \mathbb{R}^m, y \in \mathbb{R}^n : \|x\|_2 = \|y\|_2 = 1} x^\top A y$ denote the spectral norm of A, and $\|A\|_{\infty \to 1} := \max_{x \in \{-1,1\}^m, y \in \{-1,1\}^n :} x^\top A y$ denote the $(\infty \to 1)$-norm of A. We note that $\|A\|_{\infty \to 1} \leq \sqrt{nm}\|A\|_2$.

2.2 Locally decodable codes and hypergraphs

Definition 2.1. A hypergraph $\mathcal{H}$ with vertices $[n]$ is a collection of subsets $C \subseteq [n]$ called hyperedges. We say that a hypergraph $\mathcal{H}$ is *q-uniform* if $|C| = q$ for all $C \in \mathcal{H}$, and we say that $\mathcal{H}$ is a *matching* if all the hyperedges in $\mathcal{H}$ are disjoint. For a subset $Q \subseteq [n]$, we define the degree of Q in $\mathcal{H}$, denoted $\deg_{\mathcal{H}}(Q)$, to be $|\{C \in \mathcal{H} : Q \subseteq C\}|$.

Definition 2.2 (Locally Decodable Code)**.** A code $C \colon \{0,1\}^k \to \{0,1\}^n$ is (q, δ, ε)-locally decodable if there exists a randomized decoding algorithm $\mathrm{Dec}(\cdot)$ with the following properties. The algorithm $\mathrm{Dec}(\cdot)$ is given oracle access to some $y \in \{0,1\}^n$, takes an $i \in [k]$ as input, and satisfies the following: (1) the algorithm Dec makes at most q queries to the string y, and (2) for all $b \in \{0,1\}^k$, $i \in [k]$, and all $y \in \{0,1\}^n$ such that $\Delta(y, C(b)) \leq \delta$, $\Pr[\mathrm{Dec}^y(i) = b_i] \geq \frac{1}{2} + \varepsilon$. Here, $\Delta(x, y)$ denotes the relative Hamming distance between x and y, i.e., the fraction of indices $v \in [n]$ where $x_v \neq y_v$.

Following known reductions [Yek12], locally decodable codes can be reduced to the following normal form, which is more convenient to work with.

Definition 2.3 (Normal LDC)**.** A code $C \colon \{-1,1\}^k \to \{-1,1\}^n$ is (q, δ, ε)-normally decodable if for each $i \in [k]$, there is a q-uniform hypergraph matching $\mathcal{H}_i$ with at least δn hyperedges such that for every $C \in \mathcal{H}_i$, it holds that $\Pr_{b \leftarrow \{-1,1\}^k}[b_i = \prod_{v \in C} C(b)_v] \geq \frac{1}{2} + \varepsilon$.

Fact 2.4 (Reduction to LDC Normal Form, Lemma 6.2 in [Yek12])**.** *Let $C \colon \{0,1\}^k \to \{0,1\}^n$ be a code that is (q, δ, ε)-locally decodable. Then, there is a code $C' \colon \{-1,1\}^k \to \{-1,1\}^{O(n)}$ that is $(q, \delta', \varepsilon')$ normally decodable, with $\delta' \geq \varepsilon\delta/3q^2 2^{q-1}$ and $\varepsilon' \geq \varepsilon/2^{2q}$.*

2.3 Concentration inequalities

Our work will rely on the following concentration inequalities. The first is the expectation form of the standard rectangular Matrix Bernstein inequality.

Fact 2.5 (Expectation Form of Rectangular Matrix Bernstein, Theorem 1.6.2 of [Tro15])**.** *Let $X_1, \ldots, X_k$ be independent random $d_1 \times d_2$ matrices with $\mathbb{E}[X_i] = 0$ and $\|X_i\| \leq R$ for all i. Let $\sigma^2 \geq \max(\|\mathbb{E}[\sum_{i=1}^k X_i X_i^\top]\|, \|\mathbb{E}[\sum_{i=1}^k X_i^\top X_i]\|)$. Then, $\mathbb{E}[\|\sum_{i=1}^k X_i\|] \leq O(R \log(d_1 + d_2) + \sqrt{\sigma^2 \log(d_1 + d_2)})$.*

The second concentration inequality is a result for combinatorial polynomials due to Schudy and Sviridenko [SS12] that is the culmination of an influential line of work begun by Kim and Vu [KV00].

Fact 2.6 (Concentration of polynomials, Theorem 1.2 of [SS12], specialized)**.** *Let $\mathcal{H} \subseteq \binom{[n]}{t}$ be a collection of multilinear monomials of degree t in n $\{0,1\}$-valued variables, and let $f(x) \coloneqq \sum_{C \in \mathcal{H}} \prod_{i \in C} x_i$. Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed Bernoulli random variables with $\Pr[Y_i = 1] = \tau$. Then, for some absolute constant $R \geq 1$,*

$$\Pr[|f(Y) - \mathbb{E}f(Y)| \geq \lambda] \leq e^2 \max\{\max_{r=1,2,\ldots,t} e^{-\lambda^2/\nu_0 \nu_r R^t}, \max_{r=1,2,\ldots,t} e^{-(\frac{\lambda}{\nu_r R^t})^{1/r}}\},$$

where, for every $0 \leq r \leq t$, $\nu_r = \tau^{t-r} \max_{Q \subseteq [n], |Q|=r} |\{C \in \mathcal{H} : Q \subseteq C\}|$.

In this work, we will instead be working with a more convenient form of the Schudy-Sviridenko inequality, which we also prove. For our 3-query lower bound, we only require the graph version of this statement, which gives a tail bound for the number of edges that fall within a random ℓ-sized subset of vertices.

Corollary 2.7 (Fact 2.6 for sets of size ℓ)**.** *Let $\mathcal{H} \subseteq \binom{[n]}{t}$ be a t-uniform hypergraph. Let $\ell \leq n$ be an integer, and let S be drawn uniformly at random from $\binom{[n]}{\ell}$. For $r = 0, \ldots, t$, let $\nu_r = (\ell/n)^{t-r} \max_{Q \subseteq [n], |Q|=r} \deg_{\mathcal{H}}(Q)$, where $\deg_{\mathcal{H}}(Q) = |\{C \in \mathcal{H} : Q \subseteq C\}|$, and let $\nu = \max_{r=0,\ldots,t} \nu_r$. Then, for any constant $d \geq 0$, there is an absolute constant c such that*

$$\Pr_S[|\{C \in \mathcal{H} : C \subseteq S\}| \geq c\nu \log_2^t n] \leq 1/n^d \ .$$

Proof. Define the polynomial $f(x_1, \ldots, x_n) \coloneqq \sum_{C \in \mathcal{H}} x_C$. For each $S \subseteq [n]$, we associate S with the vector $x^{(S)}$ where $x_i^{(S)} = 1$ if $i \in S$, and is 0 otherwise. We note that $f(x^{(S)}) = |\{C \in \mathcal{H} : C \subseteq S\}|$. Hence, it suffices to show that $\Pr_{S \in \binom{[n]}{\ell}}[f(x^{(S)}) \geq (\nu + 1) \cdot \mathrm{polylog}(n)] \leq 1/\mathrm{poly}(n)$.

Let $\mathcal{D}$ be the distribution over $T \subseteq [n]$ where each i is added to T with probability $p = \frac{\ell}{n}\beta$ independently, and $\beta = 8(d+1)\ln n$. We will show that for any λ, it holds that

$$\Pr_{S \leftarrow \binom{[n]}{\ell}}[f(x^{(S)}) \geq \lambda] \leq \Pr_{T \leftarrow \mathcal{D}}[f(x^{(T)}) \geq \lambda] + 1/n^{d+1} \tag{1}$$

This implies Corollary 2.7. Indeed, we use Fact 2.6 to bound $\Pr_{T \leftarrow \mathcal{D}}[f(x^{(T)}) \geq \lambda]$. The parameters of interest are $\nu'_r \coloneqq p^{t-r} \max_{|Q|=r} \deg(Q) = \beta^{t-r}\nu_r \leq \beta^t \nu$. Hence, by Fact 2.6, using that $\mathbb{E}_{T \leftarrow \mathcal{D}}[f(x^{(T)})] = \nu'_0 \leq \beta^t \nu$, we conclude that there is an absolute constant c such that

$$\Pr_{T \leftarrow \mathcal{D}}[f(x^{(T)}) \geq c\beta^t \nu \log^t n] \leq 1/n^{d+1} \ .$$

It thus remains to prove Eq. (1). To do this, we will couple the two distributions as follows. First, sample $T \leftarrow \mathcal{D}$ conditioned on $|T| \geq \ell$, and then choose S to be a uniformly random subset of T of size exactly ℓ. Let $\mathcal{D}'$ be the joint distribution over pairs (S, T) with this coupling, and observe that

the induced distribution on S is simply the uniform distribution on $\binom{[n]}{\ell}$. By Chernoff bound, we have that for every $\delta \in [0,1]$,

$$\Pr_{T \leftarrow \mathcal{D}}[|T| < (1-\delta)\beta\ell] \leq \exp\left(\frac{\delta^2 \ell \beta}{2}\right).$$

Setting $\delta = 1/2$, we have that

$$\Pr_{T \leftarrow \mathcal{D}}[|T| < \ell] \leq \Pr_{T \leftarrow \mathcal{D}}[|T| < \frac{\beta\ell}{2}] \leq \exp\left(\frac{\ell\beta}{8}\right) \leq e^{\frac{\beta}{8}} \leq 1/n^{d+1}.$$

We also observe that $f(x^{(T)}) \geq f(x^{(S)})$ for any $S \subseteq T$. In particular, if $f(x^{(S)}) \geq \lambda$, then $f(x^{(T)}) \geq \lambda$. We thus have

$$\begin{aligned}\Pr_{S \leftarrow \binom{[n]}{\ell}}[f(x^{(S)}) \geq \lambda] = \Pr_{(S,T)\sim\mathcal{D}'}[f(x^{(S)}) \geq \lambda \mid |T| \geq \ell] &\leq \Pr_{(S,T)\sim\mathcal{D}'}[f(x^{(T)}) \geq \lambda \mid |T| \geq \ell] \\ &\leq \Pr_{T \leftarrow \mathcal{D}'}[f(x^{(T)}) \geq \lambda] + 1/n^{d+1},\end{aligned}$$

which proves Eq. (1) and finishes the proof. □

3 Lower Bound for 3-Query Locally Decodable Codes

In this section, we will prove Theorem 1, our main result.

By Fact 2.4, it suffices for us to show that for any code $C\colon \{-1,1\}^k \to \{-1,1\}^n$ that is $(3,\delta,\varepsilon)$-normally decodable, it holds that $k^3 \leq n \cdot \frac{O(\log^{14} n)}{\varepsilon^{16}\delta^{16}}$. As C is $(3,\delta,\varepsilon)$-normally decodable, this implies that there are 3-uniform hypergraph matchings $\mathcal{H}_1, \ldots, \mathcal{H}_k$ satisfying the property in Definition 2.3. Let $m \coloneqq \sum_{i=1}^k |\mathcal{H}_i|$ be the total number of hyperedges in the hypergraph $\mathcal{H} \coloneqq \cup_{i=1}^k \mathcal{H}_i$.

The key idea in our proof is to define a 3-XOR instance corresponding to the decoder in Definition 2.3. By Definition 2.3, the 3-XOR instance will have high value, i.e., there is an assignment to the variables satisfying a nontrivial fraction of the constraints. To finish the proof, we show that if $n \ll k^3$, then the 3-XOR instance must have small value, which is a contradiction.

We define the 3-XOR instances below.

The Key 3-XOR Instances

For each $b \in \{-1,1\}^k$, we define the 3-XOR instance Ψ_b, where:

(1) The variables are $x_1, \ldots, x_n \in \{-1,1\}$,

(2) The constraints are, for each $i \in [k]$ and $C \in \mathcal{H}_i$, $\prod_{v \in C} x_v = b_i$.

The value of Ψ_b, denoted $\mathrm{val}(\Psi_b)$, is the maximum fraction of constraints satisfied by any assignment $x \in \{-1,1\}^n$.

We associate an instance Ψ_b with the polynomial $\psi_b(x) \coloneqq \frac{1}{m}\sum_{i=1}^k b_i \sum_{C \in \mathcal{H}_i} \prod_{v \in C} x_v$, and

define $\mathrm{val}(\psi_b) := \max_{x\in\{-1,1\}^n} \psi_b(x)$. We note that $\mathrm{val}(\Psi_b) = \frac{1}{2} + \frac{1}{2}\mathrm{val}(\psi_b)$.

The first observation is that the properties of Definition 2.3 imply that the 3-XOR instances Ψ_b defined above must have reasonably large value. Formally, we have that

$$\mathbb{E}_{b\leftarrow\{-1,1\}^k}[\mathrm{val}(\psi_b)] \geq \mathbb{E}_{b\leftarrow\{-1,1\}^k}[\psi_b(C(b))] \geq 2\varepsilon \ , \tag{2}$$

where the first inequality is by definition of $\mathrm{val}(\cdot)$, and the second inequality uses Definition 2.3, as for each constraint $C \in \mathcal{H}_i$ for some i, the encoding $C(b)$ of b satisfies this constraint with probability $\frac{1}{2} + \varepsilon$ for a random b.

Overview: refuting the XOR instances. To finish the proof, it thus suffices to argue that $\mathbb{E}_{b\leftarrow\{-1,1\}^k}[\mathrm{val}(\psi_b)]$ is small. We will do this by using a CSP refutation algorithm inspired by [GKM22]. Our argument proceeds in two steps. First, we take any pair $Q = \{u, v\}$ of vertices that appears in $\gg \log n$ of the hyperedges in $\mathcal{H} := \cup_{i=1}^k \mathcal{H}_i$, and we replace this pair with a new variable y_Q in all the constraints containing this pair. This process decomposes the 3-XOR instance into a bipartite 2-XOR instance ([AGK21, GKM22]), and the residual 3-XOR instance. We then refute the bipartite 2-XOR instance, showing that its expected value is small, and then we refute the residual 3-XOR instance, where now any pair of variables appears in at most $O(\log n)$ constraints in the 3-XOR instance. Combining, we conclude that $\mathbb{E}_{b\leftarrow\{-1,1\}^k}[\mathrm{val}(\psi_b)]$ is also small, which finishes the proof.

We now formally define the decomposition process. We recall the definition of degree in hypergraphs.

Definition 3.1 (Degree)**.** Let $\mathcal{H}$ be a q-uniform hypergraph on n vertices, and let $Q \subseteq [n]$. The degree of Q, $\deg_{\mathcal{H}}(Q)$, is the number of $C \in \mathcal{H}$ with $Q \subseteq C$.

Lemma 3.2 (Hypergraph Decomposition)**.** *Let $\mathcal{H}_1, \ldots, \mathcal{H}_k$ be 3-uniform hypergraphs on n vertices, and let $\mathcal{H} := \cup_{i=1}^k \mathcal{H}_i$. Let $d \in \mathbb{N}$ be a threshold. Let $P := \{\{u, v\} : \deg_{\mathcal{H}}(\{u, v\}) > d\}$. Then, there are 3-uniform hypergraphs $\mathcal{H}'_1, \ldots, \mathcal{H}'_k$ and bipartite graphs $G_1, \ldots, G_k$, with the following properties.*

(1) Each G_i is a bipartite graph with left vertices $[n]$ and right vertices P.

(2) Each $\mathcal{H}'_i$ is a subset of $\mathcal{H}_i$.

(3) For each $i \in [k]$, there is a one-to-one correspondence between hyperedges $C \in \mathcal{H}_i \setminus \mathcal{H}'_i$ and edges e in G_i, given by $e = (w, \{u, v\}) \mapsto C = \{u, v, w\}$.

(4) Let $\mathcal{H}' := \cup_{i=1}^k \mathcal{H}'_i$. Then, for any $u \neq v \in [n]$, it holds that $\deg_{\mathcal{H}'}(\{u, v\}) \leq d$.

(5) If $\mathcal{H}_i$ is a matching, then $\mathcal{H}'_i$ and G_i are also matchings.

The proof of Lemma 3.2 is simple, and is given in Section 3.1.

Given the decomposition, the two main steps in our refutation are captured in the following two lemmas, which handle the 2-XOR and 3-XOR instances, respectively.

Lemma 3.3 (2-XOR refutation)**.** *Let P be a set, and let $G_1, \ldots, G_k$ be bipartite matchings with left vertices $[n]$ and right vertices P, where $k \geq \log_2 n$ and $|P| \leq nk/d$ for some $d \in \mathbb{N}$. For $b \in \{-1,1\}^k$, let $g_b(x,y)$ be the polynomial defined as $g_b(x,y) \coloneqq \sum_{i=1}^k b_i \sum_{e=\{v,p\}: v\in[n], p\in P} x_v y_p$, and let $\mathrm{val}(g_b) \coloneqq \max_{x\in\{-1,1\}^n, y\in\{-1,1\}^P} g_b(x,y)$. Then, $\mathbb{E}_{b\leftarrow\{-1,1\}^k}\mathrm{val}(g_b) \leq O(nk\sqrt{(\log n)/d})$.*

Lemma 3.4 (3-XOR refutation)**.** *Let $\mathcal{H}'_1, \ldots, \mathcal{H}'_k$ be 3-uniform hypergraph matchings on n vertices, and let $\mathcal{H}' \coloneqq \cup_{i=1}^k \mathcal{H}'_i$. Suppose that for any $\{u,v\} \subseteq [n]$, $\deg_{\mathcal{H}}(\{u,v\}) \leq d$, where $d = O((\log n)/\varepsilon^2\delta^2)$. Let $f_b(x) \coloneqq \sum_{i=1}^k b_i \sum_{C\in\mathcal{H}'_i} \prod_{v\in C} x_v$. Then, it holds that*

$$\mathbb{E}_{b\leftarrow\{-1,1\}^k}\mathrm{val}(f_b) \leq n\sqrt{k}\cdot O((\log^{3/2} n)/\varepsilon\delta)\cdot\left(\left(\frac{n}{k}\right)^{1/4}\sqrt{\log n} + (nk)^{1/8}\log^{1/4} n\right) \ .$$

We prove Lemma 3.3 in Section 3.2, and we prove Lemma 3.4 in Section 4.

With the above ingredients, we can now finish the proof of Theorem 1.

Proof of Theorem 1. Applying Lemma 3.2 with $d = O((\log n)/\varepsilon^2\delta^2)$ for a sufficiently large constant, we decompose the instance Ψ_b into 2-XOR and 3-XOR subinstances.[4] Note that as $m \leq nk$, we will have $|P| \leq m/d \leq nk/d$. We have that $m\mathrm{val}(\psi_b) \leq \mathrm{val}(f_b) + \mathrm{val}(g_b)$ because of the one-to-one correspondence property in Lemma 3.2. We also note that $m \geq \delta nk$, as $|\mathcal{H}_i| \geq \delta n$ for each i. By Lemma 3.3 and by taking the constant in the choice of d sufficiently large, we can ensure that $\mathbb{E}_{b\leftarrow\{-1,1\}^k}[\mathrm{val}(g_b)] \leq \varepsilon\delta nk/3$. Hence, by Eq. (2) and Lemma 3.4, we have

$$\begin{aligned}
&2\varepsilon\delta nk \leq 2\varepsilon m \leq m\mathbb{E}_{b\leftarrow\{-1,1\}^k}[\mathrm{val}(\psi_b)] \leq \mathbb{E}_{b\leftarrow\{-1,1\}^k}[\mathrm{val}(f_b) + \mathrm{val}(g_b)] \\
&\leq \frac{\varepsilon\delta nk}{3} + n\sqrt{k}\cdot O((\log^{3/2} n)/\varepsilon\delta)\cdot\left(\left(\frac{n}{k}\right)^{1/4}\sqrt{\log n} + (nk)^{1/8}\log^{1/4} n\right) \\
&\implies \varepsilon^2\delta^2\sqrt{k} \leq O(\log^{3/2} n)\cdot\left(\left(\frac{n}{k}\right)^{1/4}\sqrt{\log n} + (nk)^{1/8}\log^{1/4} n\right) \\
&\implies \varepsilon^4\delta^4 \leq O(\log^3 n)\cdot\left(\left(\frac{n}{k^3}\right)^{1/2}\log n + \left(\frac{n}{k^3}\right)^{1/4}\log^{1/2} n\right) \ .
\end{aligned}$$

There are now two cases. If $\left(\frac{n}{k^3}\right)^{1/2}\log_2 n \geq \left(\frac{n}{k^3}\right)^{1/4}\log_2^{1/2} n$, then we have that

$$\varepsilon^4\delta^4/\log_2^3 n \leq O\left(\left(\frac{n}{k^3}\right)^{1/2}\log n\right) \implies k^3 \leq n\cdot O(\log^8 n)/\varepsilon^8\delta^8 \ ,$$

and if $\left(\frac{n}{k^3}\right)^{1/2}\log_2 n \leq \left(\frac{n}{k^3}\right)^{1/4}\log_2^{1/2} n$, then we conclude that

$$\varepsilon^4\delta^4/\log_2^3 n \leq O\left(\left(\frac{n}{k^3}\right)^{1/4}\log^{1/2} n\right) \implies k^3 \leq n\cdot O(\log^{14} n)/\varepsilon^{16}\delta^{16} \ .$$

We thus conclude that $k^3 \leq n\cdot O\left(\frac{\log^{14} n}{\varepsilon^{16}\delta^{16}}\right)$, which finishes the proof. □

[4] We remark that it is possible that one (but not both!) of the 2-XOR or 3-XOR subinstances has very few constraints, or even no constraints at all. This is not a problem, however, as then the upper bound on the value of the instance shown in corresponding lemma (either Lemma 3.3 or Lemma 3.4) becomes trivial.

3.1 Hypergraph decomposition: proof of Lemma 3.2

We prove Lemma 3.2 by analyzing the following greedy algorithm.

Algorithm 3.5.

Given: 3-uniform hypergraphs $\mathcal{H}_1, \dots, \mathcal{H}_k$.

Output: 3-uniform hypergraphs $\mathcal{H}'_1, \dots, \mathcal{H}'_k$ and bipartite graphs $G_1, \dots, G_k$.

Operation:

1. **Initialize:** $\mathcal{H}'_i = \mathcal{H}_i$ for all $i \in [k]$, $P = \{\{u, v\} : \deg_{\mathcal{H}'}(\{u, v\}) > d\}$, where $\mathcal{H}' = \cup_{i \in [k]} \mathcal{H}'_i$.
2. **While** P **is nonempty**:
   (1) Choose $p = \{u, v\} \in P$ arbitrarily.
   (2) For each $i \in [k]$, $C \in \mathcal{H}'_i$ with $p \in C$, remove C from $\mathcal{H}'_i$, and add the edge $(C \setminus p, p)$ to G_i.
   (3) Recompute $P = \{\{u, v\} : \deg_{\mathcal{H}'}(\{u, v\}) > d\}$.
3. Output $\mathcal{H}'_1, \dots, \mathcal{H}'_k$, $G_1, \dots, G_k$.

Indeed, properties (1), (2) and (5) in Lemma 3.2 trivially hold. Property (4) holds because otherwise the algorithm would not have terminated, as the set P would still be nonempty. Property (3) holds because each hyperedge $C \in \mathcal{H}_i$ starts in $\mathcal{H}'_i$, and is either removed exactly once and added to G_i as $(C \setminus p, p)$, or remains in $\mathcal{H}'_i$ for the entire operation of the algorithm. This finishes the proof.

3.2 Refuting the 2-XOR instance: proof of Lemma 3.3

We now prove Lemma 3.3. We do this as follows. For each $e = \{v, p\}$, with $v \in [n]$, $p \in P$, define the matrix $A^{(e)} \in \mathbb{R}^{n \times P}$, where $A^{(e)}(v', p') = 1$ if $v' = v$ and $p' = p$, and 0 otherwise. Let $A_i \coloneqq \sum_{e \in G_i} A^{(e)}$, which is the bipartite adjacency matrix of G_i. Finally, let $A \coloneqq \sum_{i=1}^k b_i A_i$.

First, we observe that $\mathrm{val}(g_b) \le \sqrt{n|P|}\|A\|_2$. Indeed, this is because for any $x \in \{-1, 1\}^n, y \in \{-1, 1\}^P$, we have $g_b(x, y) = x^\top A y \le \|x\|_2 \|y\|_2 \|A\|_2 = \sqrt{n|P|}\|A\|_2$. Thus, in order to bound $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(g_b)]$, it suffices to bound $\mathbb{E}_b[\|A\|_2]$.

We use Fact 2.5 to bound $\mathbb{E}[\|A\|_2]$. Indeed, we observe that $\|A_i\|_2 \le 1$ for each i, as each row/column of A_i has at most one nonzero entry of magnitude 1 because each G_i is a matching. Next, we observe that $\mathbb{E}[AA^\top] = \sum_{i=1}^k A_i A_i^\top$ and $\mathbb{E}[A^\top A] = \sum_{i=1}^k A_i^\top A_i$, as the b_i's are independent, and so we conclude that $\max(\|\mathbb{E}[AA^\top]\|, \|\mathbb{E}[A^\top A]\|) \le k$. Hence, by Fact 2.5, we have that $\mathbb{E}[\|A\|_2] \le O(\log n + \sqrt{k \log n}) \le O(\sqrt{k \log n})$ where we use that $k \ge \log_2 n$. It thus follows that $\mathbb{E}[\mathrm{val}(g_b)] \le \sqrt{n|P|} O(\sqrt{k \log n}) \le O(nk\sqrt{(\log n)/d})$.

4 Refuting the 3-XOR Instance: Proof of Lemma 3.4

In this proof, we will write $\mathcal{H}_i$ instead of $\mathcal{H}'_i$ everywhere, to avoid cumbersome notation. For a vertex $u \in [n]$ and a subset $C \in \binom{[n]}{2}$, we will use the notation (u, C) to denote the set $\{u\} \cup C$.

The main idea is to construct a 4-XOR instance by "canceling" out every x_u that appears in two different clauses. Concretely, we first randomly partition $[k]$ into two sets, L, R. Then, given $(u, C_1) \in \mathcal{H}_i$ with $i \in L$ and $(u, C_2) \in \mathcal{H}_j$ with $j \in R$, we construct the derived clause $C_1 \oplus C_2$ obtained by canceling x_u. We relate the value of the derived instance to the original instance, and then produce a spectral refutation for the derived instance via an appropriate subexponential-sized matrix. This will show that the expected value of the derived instance, over the randomness of the b_i's, is small, and completes the proof.

For notation, we will let $f := f_b$, i.e., we omit the subscript, as it is clear from context. We will also let $m := |\mathcal{H}| = \sum_{i=1}^{k} |\mathcal{H}_i|$.

Relating the derived 4-XOR to the original 3-XOR. First, let (L, R) be a partition of $[k]$ into two sets. Let $f_{L,R}(x)$ be the following polynomial:

$$f_{L,R}(x) := \sum_{\substack{i \in L \\ j \in R}} \sum_{u \in [n]} \sum_{\substack{(u,C_1) \in \mathcal{H}_i \\ (u,C_2) \in \mathcal{H}_j}} b_i b_j x_{C_1} x_{C_2} \ ,$$

where x_C is defined as $\prod_{v \in C} x_v$. We note that because the $\mathcal{H}_i$'s are matchings, after fixing i, j, and u, there is at most one pair C_1, C_2 in the inner sum. As mentioned in the proof overview, the partition allows us to preserve $\sim k$ independent bits of randomness in the right hand sides of the 4-XOR instance while eliminating nontrivial correlations. This, in particular, is crucial when we eventually apply a Matrix Bernstein inequality for spectral norm of sums of independent random matrices.

The following lemma relates $\mathrm{val}(f_{L,R})$ to $\mathrm{val}(f)$.

Lemma 4.1 (Cauchy-Schwarz Trick)**.** *It holds that* $\mathrm{val}(f)^2 \le nm + 4n\mathbb{E}_{(L,R)}\mathrm{val}(f_{L,R})$. *In particular,* $\mathbb{E}_{b \in \{-1,1\}^k}\mathrm{val}(f)^2 \le nm + 4n\mathbb{E}_{(L,R)}\mathbb{E}_{b \in \{-1,1\}^k}[\mathrm{val}(f_{L,R})]$.

Proof. Fix any assignment to $x \in \{-1, 1\}^n$. We have that

$$f(x)^2 = \left(\sum_{u \in [n]} x_u \sum_{i \in [k]} \sum_{(u,C) \in \mathcal{H}_i} b_i x_C \right)^2 \le \left(\sum_{u \in [n]} x_u^2 \right) \left(\sum_{u \in [n]} \left(\sum_{i \in [k]} \sum_{(u,C) \in \mathcal{H}_i} b_i x_C \right)^2 \right)$$

$$= n \sum_{u \in [n]} \sum_{i,j \in [k]} \sum_{\substack{(u,C_1) \in \mathcal{H}_i \\ (u,C_2) \in \mathcal{H}_j}} b_i b_j x_{C_1} x_{C_2} = n \left(\sum_{i \in [k]} |\mathcal{H}_i| + \sum_{u \in [n]} \sum_{i,j \in [k], i \ne j} \sum_{\substack{(u,C_1) \in \mathcal{H}_i \\ (u,C_2) \in \mathcal{H}_j}} b_i b_j x_{C_1} x_{C_2} \right)$$

$$= nm + 4n \cdot \mathbb{E}_{(L,R)} f_{L,R}(x) \ ,$$

where the inequality follows by the Cauchy-Schwarz inequality, and the last equality follows because for a pair of hypergraphs $\mathcal{H}_i$ and $\mathcal{H}_j$, we have $i \in L$ and $j \in R$ with probability $1/4$. Hence, we have that $\mathrm{val}(f)^2 \le nm + 4n \cdot \mathbb{E}_{(L,R)}\mathrm{val}(f_{L,R})$. □

4.1 Bounding $\mathrm{val}(f_{L,R})$ using CSP refutation

It remains to bound $\mathbb{E}_{b\in\{-1,1\}^k}\mathrm{val}(f_{L,R})$ for each choice of partition (L,R). We will do this by introducing a matrix A for each $b\in\{-1,1\}^k$ and partition (L,R), and relating $\mathrm{val}_{f_{L,R}}$ to $\|A\|_{\infty\to 1}$. Note that A will depend on the choice of b and the partition (L,R). Then, we will bound $\mathbb{E}_{b\in\{-1,1\}^k}[\|A\|_{\infty\to 1}]$.

Definition 4.2. Let $C\subseteq[n]$ be a set. We let $C^{(1)}$ and $C^{(2)}$ denote the subsets of $[n]\times[2]$ defined as $C^{(b)}=\{(i,b):i\in C\}$ for $b\in[2]$, i.e., if we think of $[n]\times[2]$ as two copies of $[n]$, $C^{(1)}$ is the set C using the first copy, and $C^{(2)}$ is the set C using the second copy.

If $S\subseteq[n]\times[2]$, we will sometimes think of S as a pair (S_1,S_2), with $S_1,S_2\subseteq[n]$, and $S=S_1^{(1)}\cup S_2^{(2)}$.

Definition 4.3 (Our Kikuchi Matrix)**.** Let $\ell:=2\lceil\sqrt{n/k}\rceil$ and let $N:=\binom{2n}{\ell}$. For any two sets $S,T\subseteq[n]\times[2]$ and sets $C,C'\in\binom{[n]}{2}$, we say that $S\overset{C,C'}{\leftrightarrow}T$ if

1. $S\oplus T=C^{(1)}\oplus C'^{(2)}$,
2. $|S\cap C^{(1)}|=|S\cap C'^{(2)}|=|T\cap C^{(1)}|=|T\cap C'^{(2)}|=1$.

Note that $C^{(1)}\oplus C'^{(2)}=C^{(1)}\cup C'^{(2)}$, as $C^{(1)}$ and $C'^{(2)}$ are disjoint by construction.

For each $C,C'\in\binom{[n]}{2}$, define the $N\times N$ matrix $A^{(C,C')}$, indexed by sets $S\subseteq[n]\times[2]$ of size ℓ, by setting $A^{(C,C')}(S,T)=1$ if $S\overset{C,C'}{\leftrightarrow}T$, and 0 otherwise.

We let

$$A_{i,j}:=\sum_{u\in[n]}\sum_{(u,C)\in\mathcal{H}_i,(u,C')\in\mathcal{H}_j}A^{(C,C')},\quad A_i:=\sum_{j\in R}b_jA_{i,j},\quad\text{and}\quad A:=\sum_{i\in L}b_iA_i\ .$$

We observe that for a fixed choice of (C,C'), the matrix $A^{(C,C')}$ has exactly $D:=4\binom{2n-4}{\ell-2}$ nonzero entries. Indeed, this is the purpose of using subsets of $[n]\times[2]$ rather than just $[n]$. If we used subsets of $[n]$ only, the number of nonzero entries in $A^{(C,C')}$ would depend on $|C\oplus C'|$, whereas with subsets of $[n]\times[2]$ we always have $|C^{(1)}\oplus C'^{(2)}|=4$.

Fix an assignment $x\in\{-1,1\}^n$, and let $z\in\{-1,1\}^N$ be defined as $z_S:=\prod_{u\in S_1}x_u\prod_{v\in S_2}x_v$ for $S=(S_1,S_2)\in[n]\times[2]$ satisfying $|S|=\ell$. We observe that $Df_{L,R}(x)=z^\top Az$. Indeed, this is because for any sets $S,T\subseteq[n]\times[2]$, $z_Sz_T=\prod_{u\in S_1}x_u\prod_{v\in S_2}x_v\prod_{u'\in T_1}x_u\prod_{v'\in T_2}x_v=\prod_{u\in S_1\oplus T_1}x_u\prod_{v\in S_2\oplus T_2}x_v=\prod_{u\in C}x_u\prod_{v\in C'}x_v$. In particular, this implies

$$\mathrm{val}(f_{L,R})\le\frac{1}{D}\|A\|_{\infty\to 1}\ . \tag{3}$$

It thus remains to bound $\mathbb{E}_{b\in\{-1,1\}^k}[\|A\|_{\infty\to 1}]$.

Towards this goal, we first remove all rows/columns of A where the ℓ_1-norm of that row/column is large in some A_i, and we show that this does not appreciably affect the $(\infty\to 1)$-norm. Then, we bound the spectral norm of the "pruned" matrix, and use the spectral norm to conclude a bound on the $(\infty\to 1)$-norm. This is captured in the following two lemmas.

Definition 4.4. For a row/column S, let $\Delta_i(S)$ denote the ℓ_1-norm of the S-th row/column in A_i. Let $\Delta \coloneqq c \cdot (\log_2^3 n)/\varepsilon^2\delta^2$, where c is a sufficiently large absolute constant. Let $\mathcal{B} \coloneqq \{S : \exists i \in L, \Delta_i(S) > \Delta\}$ denote the set of "heavy" rows/columns. Let G be the matrix obtained by taking A and zero-ing out all of the heavy rows/columns, and let $B = A - G$.

The following lemma bounds the number of heavy rows/columns, and thus $\|B\|_{\infty\to 1}$.

Lemma 4.5 (Row pruning)**.** *For any $b \in \{-1,1\}^k$, we have $|\mathcal{B}| \le N/n^4$. In particular, for any choice of $b \in \{-1,1\}^k$, it holds that $\|B\|_{\infty\to 1} \le 2N/n$.*

The following lemma bounds the expected spectral norm of the matrix after removing the heavy rows/columns.

Lemma 4.6 (Spectral norm bound)**.** $\mathbb{E}_{b\in\{-1,1\}^k}[\|G\|_2] \le \Delta \cdot O(\ell \log n + \sqrt{k\ell \log n})$.

We postpone the proofs of Lemmas 4.5 and 4.6 to Sections 4.2 and 4.3, and now finish the proof.

Proof of Lemma 3.4. By Eq. (3) and Lemmas 4.5 and 4.6, we have that

$$\begin{aligned}
&\mathbb{E}_{b\in\{-1,1\}^k}[\mathrm{val}(f_{L,R})] \le \frac{1}{D}\mathbb{E}_{b\in\{-1,1\}^k}[\|A\|_{\infty\to 1}] \le \frac{1}{D}\|B\|_{\infty\to 1} + \frac{N}{D}\mathbb{E}_{b\in\{-1,1\}^k}[\|G\|_2] \\
&\le \frac{N}{D}\left(\frac{2}{n} + \Delta \cdot O(\ell \log n + \sqrt{k\ell \log n})\right) \le \frac{n^2}{\ell^2} O((\log^3 n)/\varepsilon^2\delta^2) \cdot O(\ell \log n + \sqrt{k\ell \log n}) \\
&= nk \cdot O((\log^3 n)/\varepsilon^2\delta^2) \cdot O(\sqrt{n/k}\log n + (nk)^{1/4}\sqrt{\log n}) \ ,
\end{aligned}$$

where we use that $\ell = 2\lceil\sqrt{n/k}\rceil$, $\Delta = c \cdot (\log_2^3 n)/\varepsilon^2\delta^2$, and the following simple claim.

Claim 4.7. Let $n \ge 2$, $2 \le \ell \le n$, $N = \binom{2n}{\ell}$, $D = 4\binom{2n-4}{\ell-2}$. Then, $\frac{N}{D} \le \frac{16n^2}{\ell^2}$.

Proof.

$$\begin{aligned}
&\frac{N}{D} = \frac{(2n)!}{\ell!(2n-\ell)!} \cdot \frac{(\ell-2)!(2n-\ell-2)!}{4(2n-4)!} = \frac{1}{4}\frac{(2n!)}{(2n-4)!}\frac{(\ell-2)!}{\ell!}\frac{(2n-\ell-2)!}{(2n-\ell)!} \\
&\le \frac{1}{4} \cdot (2n)^4 \cdot \frac{2}{\ell^2} \cdot \frac{2}{n^2} = \frac{16n^2}{\ell^2} \ . \quad \square
\end{aligned}$$

Finally, combining with Lemma 4.1 and using that $m \le nk$, we have that

$$\begin{aligned}
\mathbb{E}[\mathrm{val}(f)]^2 \le \mathbb{E}[\mathrm{val}(f)^2] &\le n^2k + 4n\mathbb{E}_{(L,R)}\mathbb{E}_{b\in\{-1,1\}^k}[\mathrm{val}(f_{L,R})] \\
&\le n^2k \cdot O((\log^3 n)/\varepsilon^2\delta^2) \cdot O(\sqrt{n/k}\log n + (nk)^{1/4}\sqrt{\log n}) \ .
\end{aligned}$$

Hence,

$$\mathbb{E}[\mathrm{val}(f)] \le n\sqrt{k}O((\log^{3/2} n)/\varepsilon\delta) \cdot O\left(\left(\frac{n}{k}\right)^{1/4}\sqrt{\log n} + (nk)^{1/8}\log^{1/4} n\right) \ ,$$

which finishes the proof of Lemma 3.4. $\square$

4.2 Row pruning: proof of Lemma 4.5

We will show that for a fixed $i \in L$, the number of rows with $\Delta_i(S) > \Delta$ is at most N/n^5. Lemma 4.5 then follows by union bounding over all $i \in L$, and using the fact that $|L| \leq k \leq n$.

We bound $|\mathcal{B}|$ using Corollary 2.7. Let $\mathcal{H}'$ denote the 4-uniform hypergraph with vertices $[n] \times [2]$, and hyperedges $\{C^{(1)} \oplus C'^{(2)} : \exists u \in [n], j \in R \text{ s.t. } (u, C) \in \mathcal{H}_i, (u, C') \in \mathcal{H}_j\}$. Let $\mathcal{H}''$ denote the 2-uniform hypergraph where the hyperedges are the set of all $P \subseteq [n] \times [2]$ with $|P| = 2$ such that P is contained in some hyperedge $C^{(1)} \oplus C'^{(2)}$ in $\mathcal{H}'$, and $|P \cap C^{(1)}| = |P \cap C'^{(2)}| = 1$. We include such P's with multiplicity, i.e., if a P can be defined using different choices of hyperedges in $\mathcal{H}'$, then we add P to $\mathcal{H}''$ with that multiplicity. (Note the similarity with Definition 4.3.)

Now, let ν be the parameter from Corollary 2.7 for the hypergraph $\mathcal{H}''$. Corollary 2.7 implies if $\Delta \geq O(\nu \log_2^2 n)$, then $|\{S : \Delta_i(S) > \Delta\}| \leq N/n^5$. It thus remains to argue that Δ satisfies this condition.

To do this, it suffices to compute ν. Recall from Corollary 2.7 that $\nu = \max(\nu_0, \nu_1, \nu_2)$, where

$$\nu_r := \left(\frac{\ell}{2n}\right)^{2-r} \max_{Q \subseteq [n] \times [2] : |Q| = r} \deg_{\mathcal{H}''}(Q) \ ,$$

and $\deg_{\mathcal{H}''}(Q) := |\{P \in \mathcal{H}'' : Q \subseteq P\}|$. Note that $\deg_{\mathcal{H}''}(Q) \leq O(1) \deg_{\mathcal{H}'}(Q)$, as once we fix a hyperedge $C^{(1)} \oplus C'^{(2)} \in \mathcal{H}'$, it adds at most $O(1)$ hyperedges P to $\mathcal{H}''$.

Fix r, and let us write $Q = (Q_1, Q_2)$, and let $r_1 = |Q_1|$, and $r_2 = |Q_2|$. Let

$$\nu_{r_1, r_2} = \left(\frac{\ell}{2n}\right)^{2-r_1-r_2} \max_{(Q_1, Q_2) : |Q_1| = r_1, |Q_2| = r_2} \deg_{\mathcal{H}'}(Q_1, Q_2) \ .$$

It suffices to compute ν_{r_1, r_2} for all valid choices of r_1, r_2, i.e., $0 \leq r_1, r_2 \leq 1$. We now compute ν_{r_1, r_2}.

(1) Case 1: $r_1 = r_2 = 0$. We observe that for any fixed $u \in [n]$, there are at most k hyperedges in $\cup_{j \in R} \mathcal{H}_j$ containing u, because the $\mathcal{H}_j$'s are matchings. As $\mathcal{H}_i$ is matching, each $u \in [n]$ appears in at most one $C \in \mathcal{H}_i$, and so it follows that $\deg_{\mathcal{H}'}(\emptyset) \leq nk$. So, $\nu_{0,0} \leq O(\ell^2 k/n)$.

(2) Case 2: $r_1 = 0, r_2 = 1$. Observe that for any hyperedge $C^{(1)} \oplus C'^{(2)}$ containing Q, we must have $Q_2 \subseteq C'$. It then follows that there are at most k choices for C'. This is because we must have $(u, C') \in \cup_{j \in R} \mathcal{H}_j$ for some $u \in [n]$, and there is at most one choice of C' per $j \in R$, as the $\mathcal{H}_j$'s are matchings. For each choice of C', there are at most 3 choices of $C \in \mathcal{H}_i$. Hence, we have $\deg_{\mathcal{H}'}(Q) \leq O(k)$. It follows that $\nu_{0,1} \leq O(\ell k/n)$.

(3) Case 3: $r_1 = 1, r_2 = 0$. Observe that for any hyperedge $C^{(1)} \oplus C'^{(2)}$ containing Q, we must have $Q_1 \subseteq C$. Note that we have $(u, C) \in \mathcal{H}_i$ for some $u \in [n]$, and $\mathcal{H}_i$ is a matching. Hence, there is at most one choice for C, and in particular at most 3 choices for $u \in [n]$ as well. For each choice of u, there are at most k choices of C' with $(u, C') \in \cup_{j \in R} \mathcal{H}_j$. Hence, we conclude that $\deg_{\mathcal{H}'}(Q) \leq O(k)$, and so $\nu_{1,0} \leq O(\ell k/n)$.

(4) Case 4: $r_1 = 1, r_2 = 1$. As in Case 3, there are at most 3 choices for $u \in [n]$. Now, because $Q_2 = \{v\}$ is nonempty, each C' must satisfy $v \in C'$ and $(u, C') \in \cup_{j \in R} \mathcal{H}_j$. This fixes two elements of (u, C'). Note that by assumption, the pair $\{u, v\}$ can appear in at most d constraints in $\cup_{i=1}^{k} \mathcal{H}_i$. Hence, there are at most d choices of C', and so $\deg_{\mathcal{H}'}(Q) \leq O(d)$. Hence, $\nu_{1,1} \leq O(d)$.

Now, as $\ell = \lceil\sqrt{n/k}\rceil$ and $d = O(\log n/\varepsilon^2\delta^2)$, it follows that $\nu_{0,0} \leq O(1)$, $\nu_{1,0}, \nu_{0,1} \leq O(1)$, and $\nu_{1,1} \leq O((\log n)/\varepsilon^2\delta^2)$. Hence, $\nu = O((\log n)/\varepsilon^2\delta^2)$. As $\Delta = c \cdot (\log_2^3 n)/\varepsilon^2\delta^2$, for a sufficiently large constant c, Lemma 4.5 follows.

To argue the "in particular", we first observe that $\mathcal{B}$ does not depend on b. We then note that $\|B\|_{\infty\to 1} \leq 2|\mathcal{B}| \cdot R$, where R is the maximum number of nonzero entries in a row of B. The maximum number of nonzero entries in a row is upper bounded by $nk^2 \leq n^3$, as each pair $C^{(1)} \oplus C'^{(2)}$ contributes at most one entry per row, and there are at most nk^2 of these pairs.

4.3 Spectral norm bound: proof of Lemma 4.6

Let G_i denote the matrix obtained by taking A_i and zero-ing out all heavy rows/columns. We have that $G = \sum_{i\in L} b_i G_i$ is a sum of independent, mean 0 random matrices. By construction, the ℓ_1-norm of each row/column of G_i is at most Δ. Hence, $\|G_i\|_2 \leq \Delta$. This additionally implies that $\|\sum_{i\in L} G_i G_i^\top\|_2 \leq |L|\Delta^2 \leq k\Delta^2$, and that $\|\sum_{i\in L} G_i^\top G_i\|_2 \leq |L|\Delta^2 \leq k\Delta^2$. Applying Matrix Bernstein (Fact 2.5), we conclude that $\mathbb{E}[\|G\|_2] \leq \Delta O(\log N + \sqrt{k \log N})$. As $\log N = O(\ell \log n)$, Lemma 4.6 follows.

References

[AGK21] Jackson Abascal, Venkatesan Guruswami, and Pravesh K. Kothari. Strongly refuting all semi-random boolean csps. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 454–472. SIAM, 2021.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM (JACM)*, 45(3):501–555, 1998.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of np. *Journal of the ACM (JACM)*, 45(1):70–122, 1998.

[BCG20] Arnab Bhattacharyya, L Sunil Chandran, and Suprovat Ghoshal. Combinatorial lower bounds for 3-query ldcs. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151, page 85. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2020.

[BGT17] Arnab Bhattacharyya, Sivakanth Gopi, and Avishay Tal. Lower bounds for 2-query lccs over large alphabet. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[Bri16] Jop Briët. On embeddings of ℓ_1^k from locally decodable codes. *arXiv preprint arXiv:1611.06385*, 2016.

[CGW10] Victor Chen, Elena Grigorescu, and Ronald de Wolf. Efficient and error-correcting data structures for membership and polynomial evaluation. In *27th International Symposium on Theoretical Aspects of Computer Science, STACS 2010, March 4-6, 2010, Nancy, France*, volume 5 of *LIPIcs*, pages 203–214. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2010.

[DGY11] Zeev Dvir, Parikshit Gopalan, and Sergey Yekhanin. Matching vector codes. *SIAM Journal on Computing*, 40(4):1154–1178, 2011.

[DS05] Zeev Dvir and Amir Shpilka. Locally decodable codes with 2 queries and polynomial identity testing for depth 3 circuits. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005*, pages 592–601. ACM, 2005.

[Dvi10] Zeev Dvir. On matrix rigidity and locally self-correctable codes. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity, CCC 2010, Cambridge, Massachusetts, USA, June 9-12, 2010*, pages 291–298. IEEE Computer Society, 2010.

[Efr09] Klim Efremenko. 3-query locally decodable codes of subexponential length. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 39–44. ACM, 2009.

[Fei08] Uriel Feige. Small linear dependencies for binary vectors of low weight. In *Building bridges*, volume 19 of *Bolyai Soc. Math. Stud.*, pages 283–307. Springer, Berlin, 2008.

[GKM22] Venkatesan Guruswami, Pravesh K. Kothari, and Peter Manohar. Algorithms and certificates for boolean CSP refutation: smoothed is no harder than random. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 678–689. ACM, 2022.

[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.

[Gop19] Sivakanth Gopi. Modern coding theory: lecture notes and exercises, 2019. URL: `https://homes.cs.washington.edu/~anuprao/pubs/codingtheory/exercise2.pdf`.

[IK04] Yuval Ishai and Eyal Kushilevitz. On the hardness of information-theoretic multiparty computation. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, volume 3027 of *Lecture Notes in Computer Science*, pages 439–455. Springer, 2004.

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

[KV00] Jeong Han Kim and Van H Vu. Concentration of multivariate polynomials and its applications. *Combinatorica*, 20(3):417–434, 2000.

[KW04] Iordanis Kerenidis and Ronald de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. *Journal of Computer and System Sciences*, 69(3):395–420, 2004.

[Rom06] Andrei E. Romashchenko. Reliable computations based on locally decodable codes. In *STACS 2006, 23rd Annual Symposium on Theoretical Aspects of Computer Science, Marseille, France, February 23-25, 2006, Proceedings*, volume 3884 of *Lecture Notes in Computer Science*, pages 537–548. Springer, 2006.

[SS12] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, page 437–446, USA, 2012. Society for Industrial and Applied Mathematics.

[Tre04] Luca Trevisan. Some applications of coding theory in computational complexity. *arXiv preprint cs/0409044*, 2004.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, 2015.

[WAM19] Alexander S. Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor PCA. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 1446–1468. IEEE Computer Society, 2019.

[Wol09] Ronald de Wolf. Error-correcting data structures. In *26th International Symposium on Theoretical Aspects of Computer Science, STACS 2009, February 26-28, 2009, Freiburg, Germany, Proceedings*, volume 3 of *LIPIcs*, pages 313–324. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany, 2009.

[Woo07] David Woodruff. New lower bounds for general locally decodable codes. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 14, 2007.

[Woo12] David P Woodruff. A quadratic lower bound for three-query linear locally decodable codes over any field. *Journal of Computer Science and Technology*, 27(4):678–686, 2012.

[Yek08] Sergey Yekhanin. Towards 3-query locally decodable codes of subexponential length. *Journal of the ACM (JACM)*, 55(1):1–16, 2008.

[Yek10] Sergey Yekhanin. *Locally Decodable Codes and Private Information Retrieval Schemes*. Information Security and Cryptography. Springer, 2010.

[Yek12] Sergey Yekhanin. Locally decodable codes. *Foundations and Trends in Theoretical Computer Science*, 6(3):139–255, 2012.

A CSP Refutation Proof of Existing LDC Lower Bounds

In this section, we prove the following theorem, which are the existing LDC lower bounds (up to $\mathrm{poly}(\log(n), \varepsilon, \delta)$ factors), using the connection between LDCs and CSP refutation.

Theorem A.1. *Let $C\colon \{0,1\}^k \to \{0,1\}^n$ be a code that is (q, δ, ε)-locally decodable, for constant $q \geq 2$. Then, the following hold:*

(1) If $q = 2$, $k \leq O((\log n)/\varepsilon^4\delta^2)$,

(2) If $q \neq 2$ is even, $k \leq n^{1-2/q}O((\log^{q+1} n)/\varepsilon^4\delta^2)$, and

(3) If q is odd, $k \leq n^{1-2/(q+1)}O((\log^{q+2} n)/\varepsilon^4\delta^2)$.

Proof. By Fact 2.4, it suffices to show that for a code $C\colon \{-1,1\}^k \to \{-1,1\}^n$ that is (q, δ, ε)-normally decodable, it holds that (1) $k \leq O((\log n)/\varepsilon^2\delta^2)$ if $q = 2$, (2) $k \leq n^{1-2/q}O((\log^{q+1} n)/\varepsilon^2\delta^2)$ if $q \neq 2$ is even, and (3) $k \leq n^{1-2/(q+1)}O((\log^{q+2} n)/\varepsilon^2\delta^2)$ if q is odd.

We first observe for any q, we can transform C into a code C' that is $(q+1, \delta/2, \varepsilon)$-normally decodable. In particular, it suffices to prove the lower bound in the case when q is even. We note that one can also prove the q odd case directly using a similar approach to the even case, just with asymmetric matrices. For simplicity, we do not present this proof, but the definition of the asymmetric matrices are given in Remark A.4.

Claim A.2. Let $C\colon \{-1,1\}^k \to \{-1,1\}^n$ be a code that is (q, δ, ε)-normally decodable. Then, there is a code $C'\colon \{-1,1\}^k \to \{-1,1\}^{2n}$ that is $(q+1, \delta/2, \varepsilon)$-normally decodable.

Proof. Let $C'\colon \{-1,1\}^k \to \{-1,1\}^{2n}$ be defined by setting $C'(b) = C(b)\|1^n$, i.e., the encoding of b under the original code C concatenated with n 1's. For each hypergraph $\mathcal{H}_i$, we construct the hypergraph $\mathcal{H}'_i$ as follows. First, let $\pi_i\colon \mathcal{H}_i \to [n]$ be an arbitrary ordering of the hyperedges of $\mathcal{H}_i$, and then let $\mathcal{H}'_i = \{C \cup \{n + \pi_i(C)\} : C \in \mathcal{H}_i\}$. That is, the hypergraph $\mathcal{H}'_i$ is obtained by taking each hyperedge in $\mathcal{H}_i$ and appending one of the new coordinates, and each new coordinate is added to at most one hyperedge, so that $\mathcal{H}'_i$ remains a matching. It is now obvious from construction that C' is $(q+1, \delta/2, \varepsilon)$-normally decodable, which finishes the proof. □

It thus remains to show that for any code $C\colon \{-1,1\}^k \to \{-1,1\}^n$ that is (q, δ, ε)-normally decodable with q even, it holds that $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ for $q \geq 4$ and $n \geq \exp(\Omega(k))$ for $q = 2$.

Similar to the proof of Theorem 1, we construct a q-XOR instance associated with C', and argue via CSP refutation that its value must be small. For each $b \in \{-1,1\}^k$, let Ψ_b denote the q-XOR instance with variables $x \in \{-1,1\}^n$ and constraints $\prod_{v \in C} x_v = b_i$ for all $i \in [k], C \in \mathcal{H}_i$. We let $m \coloneqq \sum_{i=1}^k |\mathcal{H}_i|$ denote the total number of constraints. Let $\psi_b(x) \coloneqq \frac{1}{m}\sum_{i=1}^k b_i \sum_{C \in \mathcal{H}_i} \prod_{v \in C} x_v$, and let $\mathrm{val}(\psi_b) \coloneqq \max_{x \in \{-1,1\}^n} \psi_b(x)$. As in the proof of Theorem 1, we observe that Definition 2.3 implies that $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\psi_b)] \geq 2\varepsilon$.

It thus remains to upper bound $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\psi_b)]$. We do this by introducing a matrix A for each $b \in \{-1,1\}^k$, where $\|A\|_{\infty \to 1}$ is related to $\mathrm{val}(\psi_b)$. We then upper bound $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\|A\|_{\infty \to 1}]$. We note that the matrix A depends on the choice of $b \in \{-1,1\}^k$ but we suppress this dependence for notational simplicity.

Definition A.3. Let $\ell := \lceil n^{1-2/q} \rceil$, and let $N := \binom{n}{\ell}$. For $C \in \binom{[n]}{q}$, we let $A^{(C)} \in \mathbb{R}^{N \times N}$ denote the matrix indexed by sets $S, T \in \binom{[n]}{\ell}$ where $A^{(C)}(S,T) = 1$ if $S \oplus T = C$, and is 0 otherwise.

We let $A_i := \sum_{C \in \mathcal{H}_i} A^{(C)}$, and $A := \sum_{i=1}^{k} b_i A_i$.

Remark A.4 (Matrices for q odd). As mentioned earlier, when q is odd we can prove the lower bound directly by choosing slightly different matrices, although we do not present the proof in full. The matrices used are defined as follows. We let the matrix $A^{(C)}$ now be indexed by rows $S \subseteq \binom{[n]}{\ell}$ and columns $T \subseteq \binom{[n]}{\ell+1}$, and let $A^{(C)}(S,T) = 1$ if $S \oplus T = C$. The matrices A_i and A are defined as before.

Our proof now proceeds as in Section 4. We similarly observe that $\mathrm{val}(\psi_b) \le \frac{1}{mD}\|A\|_{\infty \to 1}$, where $D := \binom{q}{q/2}\binom{n-q}{\ell-q/2}$ is, for a fixed $C \in \binom{[n]}{q}$, the number of pairs $(S,T) \in \binom{[n]}{\ell} \times \binom{[n]}{\ell}$ with $S \oplus T = C$, and $m := \sum_{i=1}^{k} |\mathcal{H}_i|$ is the total number of constraints. It thus remains to bound $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\|A\|_{\infty \to 1}]$.

As before, we remove all rows/columns of A where the ℓ_1-norm of that row/column is large in some A_i, and we show that this does not appreciably affect the $\infty \to 1$ norm. Then, we bound the expected spectral norm of the "pruned" matrix, and use the expected spectral norm to conclude a bound on the expected $\infty \to 1$ norm. This is captured in the following two lemmas.

Definition A.5. For a row/column S, let $\Delta_i(S)$ denote the ℓ_1-norm of the S-th row/column in A_i. Let $\Delta := c \log_2^{q/2} n$ for a large enough absolute constant c if $q \ge 3$, and $\Delta = 1$ if $q = 2$. Let $\mathcal{B} := \{S : \exists i \in L, \Delta_i(S) > \Delta\}$ denote the set of "heavy" rows/columns. Let $\mathcal{G}$ denote the complement of $\mathcal{B}$. Let G be the matrix obtained by taking A and zero-ing out all of the heavy rows/columns, and let $B = A - G$.

The following lemma bounds the number of heavy rows/columns, and thus $\|B\|_{\infty \to 1}$.

Lemma A.6 (Row pruning)**.** *For any $b \in \{-1,1\}^k$, we have $|\mathcal{B}| \le N/n^3$. In particular, for any choice of $b \in \{-1,1\}^k$, it holds that $\|B\|_{\infty \to 1} \le 2N/n$.*

The following lemma bounds the expected spectral norm of the matrix after removing the heavy rows/columns.

Lemma A.7 (Spectral norm bound)**.** $\mathbb{E}_{b \in \{-1,1\}^k}[\|G\|_2] \le \Delta O(\ell \log n + \sqrt{k\ell \log n})$.

We postpone the proofs of Lemmas A.6 and A.7 to Appendix A.1, and now finish the proof of Theorem A.1.

By Lemmas A.6 and A.7, it follows that

$$\begin{aligned}\mathbb{E}_{b \in \{-1,1\}^k}[\|A\|_{\infty \to 1}] \le \|B\|_{\infty \to 1} + N\mathbb{E}_{b \in \{-1,1\}^k}[\|G\|_2] &\le 2N/n + N\Delta O(\ell \log n + \sqrt{k\ell \log n}) \\ &= N\Delta O(\ell \log n + \sqrt{k\ell \log n}) \ .\end{aligned}$$

Hence,

$$2\varepsilon \le \mathbb{E}_{b \in \{-1,1\}^k}[\mathrm{val}(\psi_b)] \le \frac{1}{mD} N\Delta O(\ell \log n + \sqrt{k\ell \log n}) \ .$$

As $|\mathcal{H}_i| \ge \delta n$ for all i, it follows that $m \ge \delta n k$. Therefore,

$$\varepsilon \le \frac{N}{\delta n k D}\Delta O(\ell \log n + \sqrt{k\ell \log n}) \le \frac{1}{\delta n k}\left(\frac{n}{\ell}\right)^{q/2} O(\log^{q/2} n)(\ell \log n + \sqrt{k\ell \log n})$$

$$\leq \frac{1}{\delta} O(\Delta) \left(\frac{n^{1-2/q}}{k} \log n + \sqrt{\frac{n^{1-2/q}}{k} \log n} \right) \ ,$$

where we use that $\ell = \lceil n^{1-2/q} \rceil$ and also the following claim to bound $\frac{N}{D}$.

Claim A.8. Suppose that $n \geq 2\ell + q$ and $\ell \geq q - 2$. Then, $\frac{N}{D} \leq 2^q (n/\ell)^{q/2}$.

Proof.

$$\frac{N}{D} = \frac{n!}{\ell!(n-\ell)!} \cdot \frac{(\ell - q/2)!(n - \ell - q/2)!}{\binom{q}{q/2}(n-q)!} \leq \frac{n!}{(n-q)!} \cdot \frac{(\ell - q/2)!}{\ell!} \cdot \frac{(n - \ell - q/2)!}{(n-\ell)!}$$

$$\leq n^q \cdot \frac{1}{(\ell - q/2 + 1)^{q/2}} \cdot \frac{1}{(n/2)^{q/2}} \leq 2^{q/2} n^{q/2} \cdot \frac{1}{(\ell/2)^{q/2}} = 2^q \left(\frac{n}{\ell} \right)^{q/2} \ . \quad \square$$

Hence, we have shown that

$$\frac{\varepsilon \delta}{\Delta} \leq O \left(\frac{n^{1-2/q}}{k} \log n + \sqrt{\frac{n^{1-2/q}}{k} \log n} \right) \ .$$

We now have two cases. If $\frac{n^{1-2/q}}{k} \log_2 n \geq \sqrt{\frac{n^{1-2/q}}{k} \log_2 n}$, then we have

$$\frac{\varepsilon \delta}{\Delta} \leq O \left(\frac{n^{1-2/q}}{k} \log n \right) \implies k \leq n^{1-2/q} O(\Delta \log n)/\varepsilon \delta \ ,$$

and if $\frac{n^{1-2/q}}{k} \log_2 n \leq \sqrt{\frac{n^{1-2/q}}{k} \log_2 n}$, then we have

$$\frac{\varepsilon \delta}{\Delta} \leq O \left(\sqrt{\frac{n^{1-2/q}}{k} \log_2 n} \right) \implies k \leq n^{1-2/q} O(\Delta^2 \log n)/\varepsilon^2 \delta^2 \ .$$

We thus conclude that $k \leq n^{1-2/q} \cdot O(\Delta^2 \log n)/\varepsilon^2 \delta^2$. To finish the proof, we observe that for $q = 2$, we have $\Delta = 1$, and hence we must have $k \leq O((\log n)/\varepsilon^2 \delta^2)$, and for $q \geq 4$, we have $\Delta = O(\log^{q/2} n)$, and so we must have $k \leq n^{1-2/q} \cdot O((\log^{q+1} n)/\varepsilon^2 \delta^2)$. $\square$

A.1 Deferred proofs

Proof of Lemma A.6. If $q = 2$, so that $\Delta = 1$, then we have that $\ell = 1$. Let $S = \{u\}$ be a row/column. As $\mathcal{H}_i$ is a matching, it follows that u can appear in at most 1 hyperedge $C \in \mathcal{H}_i$. Thus, $\Delta_i(S) \leq 1 = \Delta$ always holds, and so $|\mathcal{B}| = 0$.

Now, suppose that $q \geq 3$, so that $\Delta := c \log^{q/2} n$, where c is a sufficiently large absolute constant, to be chosen later. We clearly have that $\Delta_i(S) = |\{C \in \mathcal{H}_i : |S \cap C| = q/2\}|$. Let $\mathcal{H}_i' := \{C' : C' \subseteq C \in \mathcal{H}_i : |C'| = q/2\}$ denote the set of half-edges from $\mathcal{H}_i$. We have that $\Delta_i(S) \leq |\{C' \in \mathcal{H}_i' : C' \subseteq S\}|$. Hence, in order to prove Lemma A.6, it suffices to argue that $\Pr_{S \leftarrow \binom{[n]}{\ell}}[|\{C' \in \mathcal{H}_i' : C' \subseteq S\}| > \Delta] \leq \frac{1}{n^3}$.

By Corollary 2.7, it thus suffices to argue that $\Delta \geq O(\nu \log^{q/2} n)$, and so we need to compute the parameter ν in Corollary 2.7. Observe that for any set Q, $\deg_{\mathcal{H}_i'}(Q) \leq O(1) \deg_{\mathcal{H}_i}(Q)$, as each hyperedge $C \in \mathcal{H}_i$ creates $\binom{q}{q/2} = O(1)$ hyperedges in $\mathcal{H}_i'$ (see Definition 3.1 for a definition of $\deg_{\mathcal{H}_i}(Q)$). Second, as $\mathcal{H}_i$ is a matching, it follows that $\deg_{\mathcal{H}_i}(Q) \leq 1$ if $Q \neq \emptyset$, and $\deg_{\mathcal{H}_i}(Q) \leq n$ if $Q = \emptyset$. Thus, we have that the parameter ν_0 is at most $\left(\frac{\ell}{n}\right)^{q/2} O(n)$, and ν_r for $1 \leq r \leq q/2$ is at most $O(1)$. As $\ell = \lceil n^{1-2/q} \rceil$, we have that $\nu_0 \leq O(1)$ also. Hence, $\nu = \max_{r=0,\ldots,q/2} \nu_r$ is $O(1)$. As $\Delta = c \log^{q/2} n$, for a sufficiently large constant c, Lemma A.6 follows.

To see the "in particular", we observe that each row in $A^{(C)}$ has at most one nonzero entry of magnitude 1. This implies that each row of A has ℓ_1-norm at most $nk \leq n^2$. Hence, $\|B\|_{\infty \to 1} \leq n^2 \cdot 2|\mathcal{B}| \leq 2N/n$. □

Proof of Lemma A.7. We will use Matrix Bernstein (Fact 2.5) to bound $\mathbb{E}[\|G\|_2]$. We write $G = \sum_{i=1}^{k} b_i G_i$, where G_i is the matrix obtained by taking A_i and zero-ing out all heavy rows/columns. We observe that $\|G_i\|_2 \leq \Delta$ by construction, as the ℓ_1-norm of any row/column of G_i is at most Δ. It then follows that $\|\mathbb{E}[G^2]\|_2 = \|\sum_{i=1}^{k} G_i^2\|_2 \leq \sum_{i=1}^{k} \|G_i\|_2^2 \leq k\Delta^2$. Hence, by Fact 2.5, it follows that $\mathbb{E}[\|G\|_2] \leq O(\Delta(\log N + \sqrt{k \log N}))$. Finally, we observe that $\log_2 N \leq \ell \log_2 n$, which finishes the proof. □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il