Electronic Colloquium on Computational Complexity, Revision 1 of 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 $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ 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 $x := \mathsf{C}(b)$ 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 $n \geq \tilde{\Omega}(k^2)$ on the blocklength.

In this paper, we prove a near-cubic lower bound of $n \geq \tilde{\Omega}(k^3)$ 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, HKM23] and, in particular, relies on bounding the spectral 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-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) $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ 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 $\Delta(y, \mathsf{C}(b)) := |\{v \in [n] : y_v \neq \mathsf{C}(b)_v\}| \leq \delta n$, 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* $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ *be a code that is* $(3, \delta, \varepsilon)$*-locally decodable. Then, it must hold that* $k^3 \leq n \cdot O((\log^6 n)/\varepsilon^{32}\delta^{16})$*. In particular, if* δ, ε *are constants, then* $n \geq \Omega(k^3/\log^6 k)$*.*

We have not attempted to optimize the dependence on ε and δ in Theorem 1; for the specific case of binary *linear* codes, one can obtain slightly better dependencies on $\log k, \varepsilon, \delta$, as we show in Theorem B.3 and Corollary B.4. It is straightforward to extend Theorem 1 to nonbinary alphabets with a polynomial loss in the alphabet size, and we do so in Theorem A.2 in Appendix A. 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] (and the refined argument in [HKM23]), which obtained strong refutation algorithms for semirandom and smoothed CSPs and proved the hypergraph Moore bound conjectured by Feige [Fei08] up to a single logarithmic factor.

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 at the end of 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, HKM23].

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 $\mathsf{C}\colon \{-1,1\}^k \to \{-1,1\}^n$. 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 $x = \mathsf{C}(b)$ 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 $x = \mathsf{C}(b)$ 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.

[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.

One might expect to show unsatisfiability of a q-XOR instance produced by a sufficiently 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 Section 5.

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 $[n] \setminus C$ 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 $x = \mathsf{C}(b)$, it is simple to observe that $\|A\|_2 \geq (\ell/n)^{q/2} \cdot \sum_{i=1}^{k} |\mathcal{H}_i| \geq (\ell/n)^{q/2} \Omega(kn)$, and this holds regardless of the draw of $b \leftarrow \{-1, 1\}^k$.

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 Khintchine, which implies that $\|A\|_2 \leq O(\Delta)(\sqrt{k\ell \log n})$ with high probability over b, 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 $\Omega(\ell)$. We show that when $\ell \leq n^{1-2/q}$, one can "zero out" rows of A_i carefully so that each row/column has at most one nonzero entry.[3] This allows us to set $\Delta = 1$ provided that $\ell \leq n^{1-2/q}$.[4]

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

[3] Concretely, one sets $A_i(S, T) = 1$ if $S \oplus T = C \in \mathcal{H}_i$, and $|S \oplus C'|, |T \oplus C'| \neq \ell$ for all other $C' \in \mathcal{H}_i \setminus C$. In other words, one sets $A_i(S, T) = 1$ if $A^{(C)}(S, T) = 1$ for some $C \in \mathcal{H}_i$ and the S-th row and T-th column are 0 in $A^{(C')}$ for all other $C' \in \mathcal{H}_i \setminus \{C\}$.

[4] The "zeroing out" step is a variant of the row pruning argument in [GKM22], which uses a sophisticated concentration

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

$$(\ell/n)^{q/2}\Omega(kn) \le \|A\|_2 \le O(\sqrt{k\ell \log n}) \ .$$

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

The case of $q = 3$. When $q = 3$, or more generally when q is odd, the matrices $A^{(C)}$ are no longer meaningful, 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 $k \le n^{1-2/(q+1)}\mathrm{polylog}(n)$, the 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 transformation 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.[5] 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 "zeroing out" step now only works if we assume that any pair $p = (u, v)$ of vertices appears in at most $\mathrm{polylog}(n)$ hyperedges in the original

inequality for polynomials [SS12] to show that almost all of the rows of A_i have ℓ_1-norm at most $\mathrm{polylog}(n)$. As shown in [HKM23], by doing this explicitly and without using concentration inequalities, we save on the $\mathrm{polylog}(n)$ factor.

[5] 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.

3-uniform hypergraph $\cup_{i=1}^{k} \mathcal{H}_i$. But, if we make this assumption, the rest of the proof follows the blueprint of the even q case, and we can prove that $n \geq k^3/\mathrm{polylog}(k)$. We note that a recent work [BCG20] managed to reprove that $n \geq k^2/\mathrm{polylog}(k)$ 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 Section 5) 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 is 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.

As we remarked above, our refutation-based proof of known q-LDC lower bounds for even q turns out to be closely related to the existing proofs [KW04, Woo07] that establish the lower bounds via a black-box reduction to 2-LDC lower bounds. Because of this, one might wonder if our lower

bound in Theorem 1 can also be proven via a black-box reduction to 2-LDC lower bounds. This turns out to be the case but only for *linear* 3-LDCs, and we present the argument in Appendix B. Curiously, our reduction-based proof requires *two* black-box invocations of the 2-LDC lower bound, which is unlike the existing proofs for even q that require only one invocation [KW04, Woo07]. Moreover, our reduction-based proof does not extend to non-linear codes; we discuss the barriers in Remark B.5.

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.

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 $\mathsf{C} \colon \{0,1\}^k \to \{0,1\}^n$ 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 $\Delta(y, \mathsf{C}(b)) \leq \delta n$, $\Pr[\mathrm{Dec}^y(i) = b_i] \geq \frac{1}{2} + \varepsilon$. Here, $\Delta(x, y)$ denotes the Hamming distance between x and y, i.e., the number 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 $\mathsf{C} \colon \{-1,1\}^k \to \{-1,1\}^n$ 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 $\Pr_{b \leftarrow \{-1,1\}^k}[b_i = \prod_{v \in C} \mathsf{C}(b)_v] \geq \frac{1}{2} + \varepsilon$.

Fact 2.4 (Reduction to LDC Normal Form, Lemma 6.2 in [Yek12])**.** *Let $\mathsf{C} \colon \{0,1\}^k \to \{0,1\}^n$ be a code that is (q, δ, ε)-locally decodable. Then, there is a code $\mathsf{C}' \colon \{-1,1\}^k \to \{-1,1\}^{O(n)}$ 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 The Matrix Khintchine inequality

Our work will use the expectation form of the standard rectangular Matrix Khintchine inequality.

Fact 2.5 (Rectangular Matrix Khintchine Inequality, Theorem 4.1.1 of [Tro15])**.** *Let $X_1, \dots, X_k$ be fixed $d_1 \times d_2$ matrices and $b_1, \dots, b_k$ be i.i.d. from $\{-1, 1\}$. Let $\sigma^2 \geq \max(\|\sum_{i=1}^k X_i X_i^\top]\|_2, \|\sum_{i=1}^k X_i^\top X_i]\|_2)$. Then*

$$\mathbb{E}\Big[\|\sum_{i=1}^k b_i X_i\|_2 \Big] \leq \sqrt{2\sigma^2 \log(d_1 + d_2)} \ .$$

2.4 A fact about binomial coefficients

We will need the following fact about the ratio of two specific binomial coefficients.

Fact 2.6. *Let n, ℓ, q be positive integers such that $n/2 \geq \ell \geq q$. Then, $e^{3q}(\ell/n)^q \geq \binom{n-2q}{\ell-q}/\binom{n}{\ell} \geq e^{-3q}(\ell/n)^q$.*

Proof. The ratio

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

This implies that

$$\binom{n-2q}{\ell-q}/\binom{n}{\ell} \leq e^{2q}\left(\frac{n-\ell}{q}\right)^q\left(\frac{\ell}{q}\right)^q \cdot 2^{-q}\left(\frac{n}{2q}\right)^{-2q} \leq e^{2q}q^{-2q}2^{-q}(2q)^{2q}\left(\frac{n-\ell}{n}\right)^q\left(\frac{\ell}{n}\right)^q \leq e^{3q}\left(\frac{\ell}{n}\right)^q \ ,$$

and that

$$\binom{n-2q}{\ell-q}/\binom{n}{\ell} \geq \left(\frac{n-\ell}{q}\right)^q\left(\frac{\ell}{q}\right)^q \cdot 2^{-2q}\left(\frac{en}{2q}\right)^{-2q} = e^{-2q} \cdot \left(\frac{n-\ell}{n}\right)^q\left(\frac{\ell}{n}\right)^q \geq e^{-2q}2^{-q}\left(\frac{\ell}{n}\right)^q \geq e^{-3q}\left(\frac{\ell}{n}\right)^q \ ,$$

where we use that $\ell \leq n/2$. Throughout, we use that $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. □

3 Lower Bound for 3-Query Locally Decodable Codes

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

Setup. By Fact 2.4, in order to show that $k^3 \leq n \cdot \frac{O(\log^6 n)}{\varepsilon^{32}\delta^{16}}$, it suffices for us to show that for any code $\mathsf{C}\colon \{-1,1\}^k \to \{-1,1\}^n$ that is $(3, \delta, \varepsilon)$-normally decodable, it holds that $k^3 \leq n \cdot \frac{O(\log^6 n)}{\varepsilon^{16}\delta^{16}}$. As C is $(3, \delta, \varepsilon)$-normally decodable, this implies that there are 3-uniform hypergraph matchings $\mathcal{H}_1, \dots, \mathcal{H}_k$ 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 we define has a 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 relevant family of 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 $\psi_b(x) := \frac{1}{m}\sum_{i=1}^{k} b_i \sum_{C \in \mathcal{H}_i} \prod_{v \in C} x_v$, and define $\mathrm{val}(\psi_b) := \max_{x \in \{-1,1\}^n} \psi_b(x)$. We note that $\mathrm{val}(\Psi_b) = \frac{1}{2} + \frac{1}{2}\mathrm{val}(\psi_b)$.

We first observe that Definition 2.3 immediately implies that every 3-XOR instance in the above family (indexed by $b \in \{-1,1\}^k$) Ψ_b must have a non-trivially large value. Formally, we have that

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

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 $\mathsf{C}(b)$ 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:

(1) **Decomposition:** 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 a residual 3-XOR instance where every pair of variables appears in at most $O(\log n)$ constraints.

(2) **Refutation:** We then produce a "strong refutation" for each of the bipartite 2-XOR and the residual 3-XOR instances that shows that the average value of the instance over the draw of $b \sim \{-1,1\}^k$ is small. This implies that each of the two instances produced and thus the original 3-XOR instance has a small expected value and finishes the proof.

We now formally define the decomposition process. We recall a notion of degree in hypergraphs that turns out to be useful in our argument (similar to the analysis in [GKM22]).

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* $\mathcal{H}' \coloneqq \cup_{i=1}^{k} \mathcal{H}_i'$. *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)**.** *Fix* $n \in \mathbb{N}$. *Let* $G_1, \ldots, G_k$ *be bipartite matchings with left vertices* $[n]$ *and a right vertex set* P *of size* $|P| \leq nk/d$ *for some* $d \in \mathbb{N}$. *For* $b \in \{-1, 1\}^k$, *let* $g_b(x, y)$ *be a homogeneous quadratic polynomial defined by*

$$g_b(x, y) \coloneqq \sum_{i=1}^{k} b_i \sum_{e=\{v,p\}: v \in [n], p \in P} x_v y_p \,,$$

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* $\{u, v\} \subseteq [n]$, $\deg_{\mathcal{H}}(\{u, v\}) \leq d$. *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*

$$\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(f_b)] \leq n\sqrt{k} \cdot O(d) \cdot (nk)^{1/8} \log^{1/4} n \quad .$$

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.[6] 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. (1) and Lemma 3.4, we have

$$\begin{aligned}
&2\varepsilon\delta nk \leq 2\varepsilon m \leq m\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\psi_b)] \leq \mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(f_b) + \mathrm{val}(g_b)] \\
&\leq \frac{\varepsilon\delta nk}{3} + n\sqrt{k} \cdot O(\sqrt{\log n}/\varepsilon\delta) \cdot (nk)^{1/8} \log^{1/4} n \\
&\implies \varepsilon^2\delta^2\sqrt{k} \leq O(\sqrt{\log n}) \cdot (nk)^{1/8} \log^{1/4} n \\
&\implies k^3 \leq n \cdot O(\log^6 n)/\varepsilon^{16}\delta^{16} \quad .
\end{aligned}$$

We thus conclude that $k^3 \leq n \cdot O\left(\frac{\log^6 n}{\varepsilon^{16}\delta^{16}}\right)$, which finishes the proof. □

[6] 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 $A_i \coloneqq \sum_{e \in G_i} A^{(e)}$, the bipartite adjacency matrix of G_i. Finally, let $A \coloneqq \sum_{i=1}^k b_i A_i$.

First, we observe that $\mathrm{val}(g_b) \leq \sqrt{n|P|}\|A\|_2$. Indeed, this is because for any $x \in \{-1, 1\}^n, y \in \{-1, 1\}^P$, we have $g_b(x, y) = x^\top A y \leq \|x\|_2 \|y\|_2 \|A\|_2 = \sqrt{n|P|}\|A\|_2$. 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 $\|A_i\|_2 \leq 1$ 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. Thus, $\max(\|\sum_{i=1}^k A_i A_i^\top\|, \|\sum_{i=1}^k A_i^\top A_i\|) \leq k$. As the b_i's are i.i.d. from $\{-1, 1\}$, by Fact 2.5 we have that $\mathbb{E}[\|A\|_2] \leq O(\sqrt{k \log n})$. It thus follows that $\mathbb{E}[\mathrm{val}(g_b)] \leq \sqrt{n|P|} O(\sqrt{k \log n}) \leq O(nk\sqrt{(\log n)/d})$.

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

In this section, we will omit the subscript and write f instead of f_b. We will also let $m \coloneqq |\mathcal{H}| = \sum_{i=1}^k |\mathcal{H}_i|$.

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 $\{u\} \cup C$. We will assume that $k \leq n/c$ for some sufficiently large absolute constant c. This is without loss of generality, as otherwise we can partition k into at most c disjoint blocks of size $\leq n/c$, and refute each of these subinstances separately.

The main idea is inspired by the "Cauchy-Schwarz" trick in the context of refuting odd-arity XOR instances. Specifically, we will construct a 4-XOR instance by "canceling" out every x_u that appears in two different clauses. Concretely, include every element in $[k]$ into one of two sets L, R uniformly at random. Then, for any $(u, C) \in \mathcal{H}_i$ with $i \in L$ and $(u, C') \in \mathcal{H}_j$ with $j \in R$, we construct the "derived clause" $C \oplus C'$ by XOR-ing both sides of the two constraints. We then relate the value of the instance with such derived constraints to the original 3-XOR instance and 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 complete the proof.

Relating the derived 4-XOR to the original 3-XOR. First, let (L, R) be a partition of $[k]$ into two sets of equal size $k/2$. Let $f_{L,R}(x)$ be the following polynomial:

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

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, C') in the inner sum. Informally speaking, only working with clauses derived across 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 is crucial in eventually applying the Matrix Khintchine inequality to produce a spectral refutation.

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

Lemma 4.1 (Cauchy-Schwarz Trick)**.** *Let f be as in Lemma 3.4 and let $L, R \subseteq [k]$ be constructed by including every element in $[k]$ to be in L with probability $1/2$ independently and defining $R = [k] \setminus L$. Then, it holds that $9 \cdot \mathrm{val}(f)^2 \leq 3nm + 4n\mathbb{E}_{(L,R)}\mathrm{val}(f_{L,R})$. In particular, $\mathbb{E}_{b \in \{-1,1\}^k}[9 \cdot \mathrm{val}(f)^2] \leq 3nm + 4n\mathbb{E}_{(L,R)}\mathbb{E}_{b \in \{-1,1\}^k}[\mathrm{val}(f_{L,R})]$.*

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

$$(3f(x))^2 = \left(\sum_{u \in [n]} x_u \sum_{i \in [k]} \sum_{(u,C) \in \mathcal{H}_i} b_i x_C \right)^2 \leq \left(\sum_{u \in [n]} x_u^2 \right) \left(\sum_{u \in [n]} \left(\sum_{i \in [k]} \sum_{(u,C) \in \mathcal{H}_i} b_i x_C \right)^2 \right)$$

$$= n \sum_{u \in [n]} \sum_{i,j \in [k]} \sum_{\substack{(u,C) \in \mathcal{H}_i \\ (u,C') \in \mathcal{H}_j}} b_i b_j x_C x_{C'} = n \left(3 \sum_{i \in [k]} |\mathcal{H}_i| + \sum_{u \in [n]} \sum_{i,j \in [k], i \neq j} \sum_{\substack{(u,C) \in \mathcal{H}_i \\ (u,C') \in \mathcal{H}_j}} b_i b_j x_C x_{C'} \right)$$

$$= 3nm + 4n \cdot \mathbb{E}_{(L,R)} f_{L,R}(x) \ ,$$

where the first equality is because there are 3 ways to decompose a set $C_i \in \mathcal{H}_i$ with $|C_i| = 3$ into a pair (u, C), 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 $1/4$. Finally, $\max_{x\in\{-1,1\}^n} \mathbb{E}_{(L,R)} f_{L,R}(x) \le \mathbb{E}_{(L,R)} \max_{x\in\{-1,1\}^n} f_{L,R}(x) = \mathbb{E}_{(L,R)}\mathrm{val}(f_{L,R})$. Thus, we have that $9 \cdot \mathrm{val}(f)^2 \le 3nm + 4n \cdot \mathbb{E}_{(L,R)}\mathrm{val}(f_{L,R})$. □

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 then we will relate $\mathrm{val}_{f_{L,R}}$ to $\|A\|_2$. Note that A will depend on the choice of b and the partition (L, R). Then, we will bound $\mathbb{E}_{b\in\{-1,1\}^k}[\|A\|_2]$.

To define the matrix A, we introduce the following definitions.

Definition 4.2. Let $u \in [n]$ be a vertex. We let $u^{(1)}$ and $u^{(2)}$ denote the elements $(u, 1)$ and $(u, 2)$ of $[n] \times [2]$, i.e., if we think of $[n] \times [2]$ as two copies of $[n]$, then $u^{(1)}$ is the first copy and $u^{(2)}$ is the second one. We use similar notation for sets, so if $C \subseteq [n]$, then $C^{(1)}$ and $C^{(2)}$ denote the subsets of $[n] \times [2]$ defined as $C^{(b)} = \{(i, b) : i \in C\}$ for $b \in [2]$.

Definition 4.3 (Half clauses)**.** For $i \in L$, $j \in R$, we define the set $P_{i,j}$ of "half clauses" to consist of all pairs $(v^{(1)}, w^{(2)})$ such that there exist clauses $(u, C) \in \mathcal{H}_i$, $(u, C') \in \mathcal{H}_j$ where $v \in C$ and $w \in C'$.

We let $P_i \coloneqq \cup_{j\in R} P_{i,j}$.

Our matrix is easiest to define in two steps. We first define a matrix B. Then, we will specify some modifications to B that yield the final matrix A.

Definition 4.4 (Our initial Kikuchi matrix)**.** Let $\ell \coloneqq (\sqrt{n/k})/c$ for some sufficiently large constant c,[7] and let $N \coloneqq \binom{2n}{\ell}$. 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 $i \in L$ and $C, C' \in \binom{[n]}{2}$, define the $N \times N$ matrix $B^{(i,C,C')}$, indexed by sets $S \subseteq [n] \times [2]$ of size ℓ, by setting $B^{(i,C,C')}(S, T) = 1$ if (1) $S \overset{C,C'}{\leftrightarrow} T$, and (2) each of S and T contains at most one half clause from P_i. Otherwise, we set $B^{(i,C,C')}(S, T) = 0$.

Finally, we let

$$B_{i,j} \coloneqq \sum_{u\in[n]} \sum_{(u,C)\in\mathcal{H}_i,(u,C')\in\mathcal{H}_j} B^{(i,C,C')}, \quad B_i \coloneqq \sum_{j\in R} b_j B_{i,j}, \quad \text{and} \quad B \coloneqq \sum_{i\in L} b_i B_i \ .$$

We note that the matrices B_i in Definition 4.4 directly give a reduction from the 3-XOR instance f to a 2-LDC, and this can be used to obtain our 3-LDC lower bound in the specific case of *linear* codes (see the proof of Theorem B.3 in Appendix B).

[7] We note that the matrix is only well-defined if $\ell \ge 2$, but this holds because we assumed that $k \le n/c'$ for some sufficiently large absolute constant c'. This is the only place where we will use this assumption.

Remark 4.5. For a fixed choice of $(u, C) \in \mathcal{H}_i$, $(u, C') \in \mathcal{H}_j$ with $j \in R$, the matrix $B^{(i,C,C')}$ has exactly $4\binom{2n-4}{\ell-2}$ nonzero entries, *if we ignore* the additional condition that S and T each contain at most one half clause from P_i. Indeed, this is because $S \overset{C,C'}{\leftrightarrow} T$ if and only if S and T each contain one entry of C and C' (2 choices per clause), and the remaining part of S and T is the same set $Q \subseteq [n] \times [2] \setminus (C^{(1)} \oplus C'^{(2)})$ of size $\ell - 2$ (which has $\binom{2n-4}{\ell-2}$ choices).

We note that this fact is the reason for using subsets of $[n] \times [2]$ rather than just $[n]$. If we used subsets of $[n]$ only, the number of nonzero entries in $B^{(i,C,C')}$ would depend on $|C \oplus C'|$, whereas with subsets of $[n] \times [2]$ we always have $|C^{(1)} \oplus C'^{(2)}| = 4$.

Observe that if $S \overset{C,C'}{\leftrightarrow} T$, then S and T each contain at least one half clause from P_i, namely coming from (C, C'). Thus, the additional condition on S and T is that they contain *no other* half clauses. As we shall show below, this additional condition implies that B_i has at most $2d$ nonzero entries per row and thus $\|B_i\|_2 \le 2d$, where d is the parameter in the statement of Lemma 3.4, *without* meaningfully affecting the number of nonzero entries in each of the $B^{(i,C,C')}$'s. We note that without this condition, one can show that $\|B_i\|_2 \ge \Omega(\ell)$, which is large.

Lemma 4.6 (Nonzero entry bound)**.** *For $i \in L$, let B_i be defined as in Definition 4.4. Then, B_i has at most $2d$ nonzero entries per row/column.*

We postpone the proof of Lemma 4.6 to Section 4.3, and now continue with the proof.

The following lemma shows that the number of nonzero entries in $B^{(i,C,C')}$ is at least $2\binom{2n-4}{\ell-2}$, i.e., half of $4\binom{2n-4}{\ell-2}$; thus, the additional condition only decreases the number of nonzero entries by a factor of 2 per derived constraint. The factor of 2 is not important and is chosen for convenience, and determines the constant c in the parameter ℓ.

Lemma 4.7 (Counting nonzero entries)**.** *For some $(u, C) \in \mathcal{H}_i$ and $(u, C') \in \mathcal{H}_j$ with $j \in R$, let $B^{(i,C,C')}$ be as in Definition 4.4. Then, the number of nonzero entries in $B^{(i,C,C')}$ is at least $2\binom{2n-4}{\ell-2}$.*

We postpone the proof of Lemma 4.7 to Section 4.2, and now continue with the proof.

We obtain the final matrix A by, for each $A^{(i,C,C')}$, zero-ing out entries of $B^{(i,C,C')}$ until it has *exactly* $2\binom{2n-4}{\ell-2}$ nonzero entries. This is identical to the "equalizing step" of the edge deletion process in [HKM23].

Definition 4.8 (Our final Kikuchi matrix)**.** For each $i \in L$ and each pair of clauses $(u, C) \in \mathcal{H}_i$ and $(u, C') \in \mathcal{H}_j$ with $j \in R$, let $A^{(i,C,C')}$ be the matrix obtained from $B^{(i,C,C')}$ by arbitrarily zero-ing out entries of $B^{(i,C,C')}$ until the resulting matrix has exactly $D \coloneqq 2\binom{2n-4}{\ell-2}$ nonzero entries.

We let

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

We are now ready to finish the proof. First, we relate $\|A\|_2$ to $\mathrm{val}(f_{L,R})$. Fix an assignment $x \in \{-1, 1\}^n$, and let $z \in \{-1, 1\}^N$ be defined as $z_S \coloneqq \prod_{u \in S_1} x_u \prod_{v \in S_2} x_v$ for $S = S_1^{(1)} \cup S_2^{(2)} \subseteq [n] \times [2]$ satisfying $|S| = \ell$.

We observe that $D \cdot f_{L,R}(x) = z^\top A z$. This is because:

(1) For $S, T \subseteq [n] \times [2]$ with $S \oplus T = C^{(1)} \oplus C'^{(2)}$, we have
$z_S z_T = \prod_{u \in S_1} x_u \prod_{v \in S_2} x_v \prod_{u' \in T_1} x_u \prod_{v' \in T_2} x_v = \prod_{u \in S_1 \oplus T_1} x_u \prod_{v \in S_2 \oplus T_2} x_v = \prod_{u \in C} x_u \prod_{v \in C'} x_v$,

(2) For a pair of clauses $(u, C) \in \mathcal{H}_i$ and $(u, C') \in \mathcal{H}_j$ with $i \in L$ and $j \in R$, there are exactly $D = 2\binom{2n-4}{\ell-2}$ nonzero entries (S, T) of $A^{(i,C,C')}$, and these entries have $S \oplus T = C^{(1)} \oplus C'^{(2)}$.

In particular, this implies

$$\mathrm{val}(f_{L,R}) \leq \frac{N}{D} \cdot \|A\|_2 \ . \tag{2}$$

It thus remains to bound $\mathbb{E}_{b \in \{-1,1\}^k}[\|A\|_2]$, which we do in the following lemma.

Lemma 4.9 (Spectral norm bound)**.** $\mathbb{E}_{b \in \{-1,1\}^k}[\|A\|_2] \leq d \cdot O(\sqrt{k\ell \log n})$.

We postpone the proof of Lemma 4.9 to Section 4.3, and now finish the proof of Lemma 3.4.

Proof of Lemma 3.4. By Eq. (2) and Lemma 4.9, we have that

$$\begin{aligned}
\mathbb{E}_{b \in \{-1,1\}^k}[\mathrm{val}(f_{L,R})] &\leq \frac{N}{D} \mathbb{E}_{b \in \{-1,1\}^k}[\|A\|_2] \\
&\leq \frac{N}{D} \left(d \cdot O(\sqrt{k\ell \log n}) \right) \leq \frac{n^2}{\ell^2} d \cdot O(\sqrt{k\ell \log n}) \\
&= nkd \cdot O((nk)^{1/4} \sqrt{\log n}) \ ,
\end{aligned}$$

where we use that $\ell = (\sqrt{n/k})/c$ for some constant c, and we use Fact 2.6 to bound N/D. Finally, combining with Lemma 4.1 and using that $m \leq nk$, we have that

$$\begin{aligned}
\mathbb{E}[\mathrm{val}(f)]^2 \leq \mathbb{E}[\mathrm{val}(f)^2] &\leq \frac{1}{9} \cdot \left(3n^2 k + 4n \mathbb{E}_{(L,R)} \mathbb{E}_{b \in \{-1,1\}^k}[\mathrm{val}(f_{L,R})] \right) \\
&\leq n^2 kd \cdot O((nk)^{1/4} \sqrt{\log n}) \ .
\end{aligned}$$

Hence,

$$\mathbb{E}[\mathrm{val}(f)] \leq n\sqrt{kd} \cdot O\left((nk)^{1/8} \log^{1/4} n \right) \ ,$$

which finishes the proof of Lemma 3.4. □

4.2 Counting nonzero entries: proof of Lemma 4.7

Proof of Lemma 4.7. Fix $j \in R$ and clauses $(u, C) \in \mathcal{H}_i$ and $(u, C') \in \mathcal{H}_j$. Recall that in Remark 4.5, we observed that there are exactly $4\binom{2n-4}{\ell-2}$ pairs (S, T) with $S \overset{C,C'}{\leftrightarrow} T$. Indeed, this is because $S \overset{C,C'}{\leftrightarrow} T$ if and only if S and T each contain one entry of C and C' (2 choices per clause), and the remaining part of S and T is the same set $Q \subseteq [n] \times [2] \setminus (C^{(1)} \oplus C'^{(2)})$ of size $\ell - 2$ (which has $\binom{2n-4}{\ell-2}$ choices).

From the above, we observe that for each $Q \subseteq [n] \times [2] \setminus (C^{(1)} \oplus C'^{(2)})$ of size $\ell - 2$, we can identify Q with 4 different pairs (S, T) with $S \overset{C,C'}{\leftrightarrow} T$; namely, each pair (S, T) corresponds to a subset of size 2 of (C, C') containing exactly one entry from each of C, C'. We note that these 4 choices of (S, T) correspond exactly to the 4 half clauses in P_i contributed by the derived clause (C, C'). We will

show that for at least $\frac{1}{2}\binom{2n-4}{\ell-2}$ choices of Q, *all* 4 corresponding choices of (S,T) will contain exactly one derived clause from P_i: namely, the half clause of (C,C') that we add to Q to obtain S or T. This clearly suffices to finish the proof.

Call such a set Q *bad* if it does not have the above property, i.e., there is some pair (S,T) identified with Q such that one of S or T contains more than one half clause from P_i. Since $S \overset{C,C'}{\leftrightarrow} T$ already implies that each of S and T has exactly one half clause from $C^{(1)} \oplus C'^{(2)}$, there are three ways that Q can be bad:

(1) Q contains a half clause from P_i,

(2) there is $v^{(1)} \in C^{(1)}$ and $w^{(2)} \in Q$ such that $(v^{(1)}, w^{(2)}) \in P_i$,

(3) there is $v^{(1)} \in Q$ and $w^{(2)} \in C'^{(2)}$ such that $(v^{(1)}, w^{(2)}) \in P_i$.

We thus have that the number of bad Q's is at most

$$p_0\binom{2n-6}{\ell-4} + p_1\binom{2n-5}{\ell-3} + p_2\binom{2n-5}{\ell-3} \ ,$$

where $p_0 = |P_i|$, $p_1 = |\{(v^{(1)}, w^{(2)}) \in P_i : v^{(1)} \in C^{(1)}\}|$, $p_2 = |\{(v^{(1)}, w^{(2)}) \in P_i : w^{(2)} \in C'^{(2)}\}|$.

We now upper bound p_0, p_1, p_2. Recall that a half clause in P_i is a pair $(v^{(1)}, w^{(2)})$ such that there are clauses $(u, C_1) \in \mathcal{H}_i$, $(u, C_2) \in \mathcal{H}_j$ with $j \in R$, and $v \in C_1$, $w \in C_2$.

(1) We have $p_0 \le 4nk$, as for each $u \in [n]$, because the $\mathcal{H}_i$'s are matchings, there is at most one C_1 such that $(u, C_1) \in \mathcal{H}_i$, and at most k choices of $(u, C_2) \in \mathcal{H}_j$ with $j \in R$, as $|R| \le k$. Finally, each choice of (C_1, C_2) yields 4 half clauses.

(2) We have $p_1 \le 8k$. First, there are at most 2 choices for v, each coming from C. For each such v, there is at most one $C_i \in \mathcal{H}_i$ with $v \in C_i$. (Note that $|C_i| = 3$.) Once C_i is fixed, we have at most 2 choices for u, given by $C_i \setminus \{v\}$, and there are at most k hyperedges $(u, C_2) \in \mathcal{H}_j$ for $j \in R$ (as each $\mathcal{H}_j$ is a matching and $|R| \le k$). Finally, for each such C_2 there are 2 possible choices for w.

(3) We have $p_2 \le 8k$. First, there are at most 2 choices for w, each coming from C'. For each such w, there are at most k choices of $C_j \in \cup_{j \in R} \mathcal{H}_j$ with $w \in C_j$, as each $\mathcal{H}_j$ is a matching and $|R| \le k$. (Note that $|C_j| = 3$.) For each such C_j, there are at most 2 choices for u, given by $C_j \setminus \{w\}$, and for each u, there is at most one choice of C_1 such that $(u, C_1) \in \mathcal{H}_i$. Finally, such a C_1, if it exists, gives 2 choices for v.

Combining, we thus have that the number of bad Q's is at most

$$4nk\binom{2n-6}{\ell-4} + 16k\binom{2n-5}{\ell-3} \ .$$

We have that

$$\frac{4nk\binom{2n-6}{\ell-4} + 16k\binom{2n-5}{\ell-3}}{\binom{2n-4}{\ell-2}} = \frac{4nk\frac{(2n-6)!}{(\ell-4)!(2n-2-\ell)!} + 16k\frac{(2n-5)!}{(\ell-3)!(2n-2-\ell)!}}{\frac{(2n-4)!}{(\ell-2)!(2n-2-\ell)!}}$$
$$= 4nk\frac{(\ell-2)(\ell-3)}{(2n-4)(2n-5)} + 16k\frac{\ell-2}{2n-4} \le \frac{1}{2} \ ,$$

as we have $\ell \le (\sqrt{n/k})/c$, for some sufficiently large constant c, and $k \le \sqrt{nk}$ since $k \le n$. □

4.3 Spectral norm bound: proof of Lemmas 4.6 and 4.9

Proof of Lemma 4.6. Fix $i \in L$. We show that each row/column of B_i has at most $2d$ nonzero entries. Indeed, this is because if S is a nonzero row (or column) in B_i, then S contains at most one half clause from P_i. If (C, C') is a derived clause where $S \overset{C,C'}{\leftrightarrow} T$ for some T, then S must contain a half clause in P_i that is contained in $C^{(1)} \oplus C'^{(2)}$, i.e., a half clause coming from (C, C'). As S contains at most one half clause, it follows that the number of nonzero entries in the S-th row is upper bounded by the maximum, over all half clauses, of the number of derived clauses (C, C') that contain this half clause. One can observe that this is $2d$. Indeed, if we fix $v^{(1)}$ and $w^{(2)}$, there is at most one clause $C \in \mathcal{H}_i$ containing v. Once v is fixed, there are two choices for u in $C \setminus \{v\}$. Once we have chosen u, the second clause must be $(u, C') \in \mathcal{H}_j$ for some $j \in R$, where C' contains w. By assumption, the number of hyperedges in $\cup_{i=1}^{k} \mathcal{H}_i$ containing the pair $\{u, w\}$ is at most d, so there are at most d choices for C'. □

Proof of Lemma 4.9. We have that $A = \sum_{i \in L} b_i A_i$, where the b_i's are i.i.d. from $\{-1, 1\}$. By Lemma 4.6, we know that the number of nonzero entries in a row/column of B_i is at most $2d$. As A_i is obtained by zero-ing out entries of B_i, it follows that this also holds for A_i. It thus follows that the ℓ_1-norm of any row/column of A_i is at most $2d$, and thus $\|A_i\|_2 \leq 2d$. This additionally implies that $\|\sum_{i \in L} A_i A_i^\top\|_2 \leq |L|(2d)^2 \leq k(2d)^2$, and that $\|\sum_{i \in L} A_i^\top A_i\|_2 \leq |L|(2d)^2 \leq k(2d)^2$. Applying Matrix Khintchine (Fact 2.5), we conclude that $\mathbb{E}[\|A\|_2] \leq d \cdot O(\sqrt{k \log N})$. As $\log N = O(\ell \log n)$, Lemma 4.9 follows. □

5 CSP Refutation Proof of Existing LDC Lower Bounds

In this section, we prove the following theorem, which are the existing LDC lower bounds using the connection between LDCs and CSP refutation.

Theorem 5.1. *Let $\mathsf{C} \colon \{0,1\}^k \to \{0,1\}^n$ be a code that is (q, δ, ε)-locally decodable, for constant $q \geq 2$. Then, the following hold:*

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

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

Proof. By Fact 2.4, it suffices to show that for a code $\mathsf{C} \colon \{-1,1\}^k \to \{-1,1\}^n$ that is (q, δ, ε)-normally decodable, it holds that (1) $k \leq n^{1-2/q} O((\log n)/\varepsilon^2\delta^2)$ if q is even, and (2) $k \leq n^{1-2/(q+1)} O((\log n)/\varepsilon^2\delta^2)$ 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 is given in Remark 5.4.

Claim 5.2. Let $\mathsf{C} \colon \{-1,1\}^k \to \{-1,1\}^n$ be a code that is (q, δ, ε)-normally decodable. Then, there is a code $\mathsf{C}' \colon \{-1,1\}^k \to \{-1,1\}^{2n}$ that is $(q+1, \delta/2, \varepsilon)$-normally decodable.

Proof. Let $\mathsf{C}' \colon \{-1,1\}^k \to \{-1,1\}^{2n}$ be defined by setting $\mathsf{C}'(b) = \mathsf{C}(b)\|1^n$, 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 $\mathsf{C} \colon \{-1,1\}^k \to \{-1,1\}^n$ 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 $q = 2$. Without loss of generality, we may assume that the hypergraphs $\mathcal{H}_1, \ldots, \mathcal{H}_k$ all have size *exactly* δn.

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 $\|A\|_2$ is related to $\mathrm{val}(\psi_b)$. We then upper bound $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\|A\|_2]$. 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 5.3. Let $\ell \coloneqq n^{1-2/q}/c$ for some absolute constant $c \geq e^{16}$, and let $N \coloneqq \binom{n}{\ell}$. For each q-uniform hypergraph matching $\mathcal{H}_i$, let $A_i \in \mathbb{R}^{N \times N}$ denote the matrix indexed by sets $S, T \in \binom{[n]}{\ell}$ where $A_i(S,T) = 1$ if the pair (S,T) satisfies (1) $S \oplus T = C \in \mathcal{H}_i$, and (2) $|S \oplus C'| \neq \ell$, $|T \oplus C'| \neq \ell$ for every $C' \in \mathcal{H}_i$ with $C' \neq C$. We set $A_i(S,T) = 0$ otherwise. We let $A \coloneqq \sum_{i=1}^k b_i A_i$.

Remark 5.4 (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_i now be indexed by rows $S \in \binom{[n]}{\ell}$ and columns $T \in \binom{[n]}{\ell+1}$, and let $A_i(S,T) = 1$ if $S \oplus T = C \in \mathcal{H}_i$, and $|S \oplus C'| \neq \ell + 1, |T \oplus C'| \neq \ell$, for all $C' \in \mathcal{H}_i$ with $C' \neq C$. The matrix A is again defined as $\sum_{i=1}^k b_i A_i$.

Lemma 5.5. *There is an integer D such that the following holds. Fix $i \in [k]$, and let A_i be one of the matrices defined in Definition 5.3. For any $C \in \mathcal{H}_i$, the number of pairs (S,T) with $S \oplus T = C$ and $A_i(S,T) = 1$ is exactly D. Moreover, we have that $D/N \geq \frac{1}{2}\binom{q}{q/2} e^{-3q} \cdot (\frac{\ell}{n})^{q/2}$.*

We postpone the proof of Lemma 5.5, and now finish the proof.

Our proof now proceeds as in Section 4. We similarly observe that $\mathrm{val}(\psi_b) \leq \frac{N}{mD}\|A\|_2$, where D is from Lemma 5.5, and $m \coloneqq \sum_{i=1}^k |\mathcal{H}_i|$ is the total number of constraints. It thus remains to bound $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\|A\|_2]$, which we do in the following lemma.

Lemma 5.6 (Spectral norm bound)**.** $\mathbb{E}_{b \in \{-1,1\}^k}[\|A\|_2] \leq O(\sqrt{k\ell \log n})$.

Proof. We will use Matrix Khintchine (Fact 2.5) to bound $\mathbb{E}[\|A\|_2]$. We have $A = \sum_{i=1}^{k} b_i A_i$. We observe that $\|A_i\|_2 \leq 1$ by construction, as the ℓ_1-norm of any row/column of A_i is at most 1. It then follows that $\|\sum_{i=1}^{k} A_i^2\|_2 \leq \sum_{i=1}^{k} \|A_i\|_2^2 \leq k$. Hence, by Fact 2.5, it follows that $\mathbb{E}[\|A\|_2] \leq O(\sqrt{k \log N})$. Finally, we observe that $\log_2 N \leq \ell \log_2 n$, which finishes the proof. □

We now finish the proof of Theorem 5.1. By Lemma 5.6, we have

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

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

$$\varepsilon \leq \frac{N}{\delta n k D} O(\sqrt{k\ell \log n}) \leq \frac{1}{\delta n k} \left(\frac{n}{\ell}\right)^{q/2} \cdot O(\sqrt{k\ell \log n}) \leq \frac{1}{\delta} \cdot O\left(\sqrt{\frac{n^{1-2/q}}{k} \log n}\right) \ ,$$

where we use that $\ell = n^{1-2/q}/c$ and the bound on $\frac{D}{N}$ from Lemma 5.5. We thus conclude that $k \leq n^{1-2/q} \cdot O(\log n)/\varepsilon^2 \delta^2$. □

Proof of Lemma 5.5. First, let $C \in \mathcal{H}_i$ be any element. We first show that the number of pairs (S, T) with $S \oplus T = C$ and $A_i(S, T) = 1$ is independent of C. Indeed, let $C' \in \mathcal{H}_i$ be different from C. As $\mathcal{H}_i$ is a matching, we have that C and C' are disjoint. Let π be an arbitrary bijection between C and C' and extend π to act on all of $[n]$ by acting as the identity on elements not in $C \cup C'$. It is simple to observe that if (S, T) is any pair satisfying the above criterion for C, then (S', T'), obtained by applying π to all elements of S and T, satisfies the criterion for C'. Hence, the number of pairs is independent of the choice of $C \in \mathcal{H}_i$.

We note that it is clear from symmetry that D depends only on $|\mathcal{H}_i|$, q, and n. As $|\mathcal{H}_i| = \delta n$ for all i, it follows that D does not depend on i.

We now lower bound D. Let $C \in \mathcal{H}_i$ be arbitrary. We observe that $S \oplus T = C$ if and only if $S = C_S \cup Q$, $T = C_T \cup Q$, where $C_S, C_T \subseteq C$ are disjoint subsets of size exactly $q/2$, so that $C = C_S \cup C_T$, $Q \subseteq [n] \setminus C$ has size exactly $\ell - q/2$. It follows that if $S \oplus T = C$ and for some $C' \neq C \in \mathcal{H}_i$, either $|S \oplus C'| = \ell$ or $|T \oplus C'| = \ell$, then it must be the case that $|Q \cap C'| = q/2$. Hence, we have that

$$D \geq \binom{q}{q/2}\binom{n-q}{\ell - q/2} - |\mathcal{H}_i| \cdot \binom{q}{q/2}^2 \binom{n-2q}{\ell - q} \ .$$

Applying Fact 2.6, we thus have that

$$\begin{aligned} D/N &\geq \binom{q}{q/2} e^{-3q} \left(\frac{\ell}{n}\right)^{q/2} - n \cdot \binom{q}{q/2}^2 e^{3q} \left(\frac{\ell}{n}\right)^{q} \\ &= \binom{q}{q/2} e^{-3q} \left(\frac{\ell}{n}\right)^{q/2} \left(1 - n \cdot 2^q e^{6q} \left(\frac{\ell}{n}\right)^{q/2}\right) \\ &\geq \frac{1}{2} \binom{q}{q/2} e^{-3q} \left(\frac{\ell}{n}\right)^{q/2} \ , \end{aligned}$$

where we use that $\ell \leq n^{1-2/q}/e^{16}$. □

Acknowledgements

We thank the anonymous reviewers for their helpful comments on an earlier draft of the paper. We also thank Tim Hsieh and Sidhanth Mohanty for helpful discussions.

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`.

[HKM23] Jun-Ting Hsieh, Pravesh K. Kothari, and Sidhanth Mohanty. A simple and sharper proof of the hypergraph moore bound. *ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2023.

[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.

[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.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[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 Improved Lower Bounds for 3-LDCs over Larger Alphabets

In this appendix, we will extend Theorem 1 to 3-query LDCs over larger alphabets, which will follow from combining Theorem 1 with standard results from [KT00, KW04]. We first define LDCs over general alphabets.

Definition A.1 (LDCs over general alphabets)**.** Given a positive integer q, constants $\delta, \varepsilon > 0$, and an alphabet Σ, we say a code $\mathsf{C}\colon \{0,1\}^k \to \Sigma^n$ is (q, δ, ε)-locally decodable code (abbreviated (q, δ, ε)-LDC) 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 $y \in \Sigma^n$, 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 $y \in \Sigma^n$ such that $\Delta(y, \mathsf{C}(b)) \le \delta n$, $\Pr[\mathrm{Dec}^y(i) = b_i] \ge \frac{1}{2} + \varepsilon$.

Our extension of Theorem 1 to larger alphabets is the following theorem.

Theorem A.2. *Let $\mathsf{C}\colon \{0,1\}^k \to \Sigma^n$ be a $(3, \delta, \varepsilon)$-LDC. Then, it must hold that $k^3 \le |\Sigma|^{41} n \cdot O(\log^6(|\Sigma|n)/\varepsilon^{32}\delta^{16})$. In particular, if δ, ε are constants and $|\Sigma| \le n$, then $n \ge \Omega(k^3/(|\Sigma|^{41} \log^6 k))$.*

To prove Theorem A.2, it suffices to show the following lemma.

Lemma A.3. *Let $\mathsf{C}\colon \{0,1\}^k \to \Sigma^n$ be a $(3, \delta, \varepsilon)$-LDC. Then, there exists a binary code $\mathsf{C}'\colon \{0,1\}^k \to \{0,1\}^{n'}$ with $n' \le 4n|\Sigma|$ and q-uniform matchings $\mathcal{H}'_1, \ldots, \mathcal{H}'_k$ over n' vertices such that for all $i \in [k]$, we have $|\mathcal{H}'_i| \ge \varepsilon\delta n'/(4q^2|\Sigma|)$. Furthermore, for any query set $C \in \mathcal{H}'_i$, we have that $\Pr_{b \leftarrow \{0,1\}^k}[b_i = \oplus_{v \in C} \mathsf{C}(b)_v] \ge \frac{1}{2} + \frac{\varepsilon}{8|\Sigma|^{3/2}}$.*

Indeed, once we have Lemma A.3, then by applying Theorem 1 on the resulting normal LDC,[8] we obtain Theorem A.2. Now, to prove Lemma A.3, we first need the following result from [KT00].

Lemma A.4 (Theorem 1 + Lemma 4 in [KT00])**.** *Let $\mathsf{C}\colon \{0,1\}^k \to \Sigma^n$ be a (q, δ, ε)-LDC. Then, there exists q-uniform matchings $\mathcal{H}_1, \ldots, \mathcal{H}_k$ over $[n]$ such that for all $i \in [k]$, we have $|\mathcal{H}_i| \ge \varepsilon\delta n/q^2$. Furthermore, for any query set $C \in \mathcal{H}_i$, there exists a function $f_C\colon \Sigma^q \to \{0,1\}$ such that $\Pr_{b \leftarrow \{0,1\}^k}[b_i = f_C(\mathsf{C}(b)|_C)] \ge \frac{1}{2} + \frac{\varepsilon}{2}$.*

Note that formally the statement in [KT00] only guarantees that each query set in $\mathcal{H}_i$ has size at most q rather than *exactly* q. However, we can trivially make each set be of size exactly q by padding each codeword of C with n zeros.

Next, we need the following lemma, which is a generalized and improved version of a similar lemma appearing in [KW04].

Lemma A.5 (Lemma 2 of [KW04])**.** *Let $q \ge 2$ be an integer and let $\mathsf{C}\colon \{0,1\}^k \to \Sigma^n$ be a code. Let $\mathcal{H}_1, \ldots, \mathcal{H}_k$ be q-uniform matchings over $[n]$ such that for each $i \in [k]$, we have $|\mathcal{H}_i| \ge \varepsilon\delta n/q^2$, and suppose that for each $C \in \mathcal{H}_i$, there exists a function $f_C\colon \Sigma^q \to \{0,1\}$ such that $\Pr_{b \leftarrow \{0,1\}^k}[b_i = f_C(\mathsf{C}(b)|_C)] \ge \frac{1}{2} + \frac{\varepsilon}{2}$.*

Then, there exists a binary code $\mathsf{C}'\colon \{0,1\}^k \to \{0,1\}^{n'}$ with $n' \le 4n|\Sigma|$ and q-uniform matchings $\mathcal{H}'_1, \ldots, \mathcal{H}'_k$ over n' vertices such that for all $i \in [k]$, we have $|\mathcal{H}'_i| \ge \varepsilon\delta n'/(4q^2|\Sigma|)$. Furthermore, for any query set $C \in \mathcal{H}'_i$, we have that $\Pr_{b \leftarrow \{0,1\}^k}[b_i = \oplus_{v \in C} \mathsf{C}'(b)_v] \ge \frac{1}{2} + \frac{\varepsilon}{2^q|\Sigma|^{q/2}}$.

Combining Lemma A.4 and Lemma A.5, we immediately obtain Lemma A.3; Theorem A.2 then follows by applying Theorem 1. Thus, it remains to prove Lemma A.5. In what follows, we use conventional notations of Boolean analysis from [O'D14].

Proof of Lemma A.5. Consider a natural number $\ell \in \mathbb{N}$ such that $|\Sigma| < 2^\ell \le 2|\Sigma|$, and let $n' := n2^{\ell+1}$. Without loss of generality, say that $\Sigma \subseteq \{0,1\}^\ell$. Consider the first-order Reed-Muller encoding $\mathrm{RM}_1\colon \{0,1\}^\ell \to \{0,1\}^{2^{\ell+1}}$ defined as $\mathrm{RM}_1(\sigma) = (\langle a, \sigma\rangle + t)_{a \in \{0,1\}^\ell, t \in \{0,1\}}$.[9] We define our new code $\mathsf{C}'\colon \{0,1\}^k \to \{0,1\}^{n'}$ as $\mathsf{C}'(b) := (\mathrm{RM}_1(\mathsf{C}(b)_1), \ldots, \mathrm{RM}_1(\mathsf{C}(b)_n))$.

[8] Note that we obtain a better dependence on ε in Theorem 1 when our initial LDC is in normal form, as shown at the beginning of Section 3.

[9] Here, $\langle \cdot, \cdot \rangle$ denotes the pointwise inner product over $\mathbb{F}_2^\ell$.

Consider any message index $i \in [k]$ and query set $C \in \mathcal{H}_i$. We are going to find a corresponding query set for C in C'. Write $C = \{v_1, \ldots, v_q\}$. Arbitrarily extend our function f_C to a function over $(\{0,1\}^\ell)^q$ by setting $f_C(\sigma) = 0$ for $\sigma \in \{0,1\}^\ell \setminus \Sigma$. For any message $b \in \{0,1\}^k$, set $x \coloneqq \mathsf{C}(b)$. Switching from $\{0,1\}$ to $\{-1,1\}$ in the natural way, we find that

$$\Pr_{b \leftarrow \{0,1\}^k}[b_i = f_C(\mathsf{C}(b)|_C)] \geq \frac{1}{2} + \frac{\varepsilon}{2} \iff \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k}[b_i f_C(x_{v_1}, \ldots, x_{v_q})] \geq \varepsilon \ .$$

Consider the Fourier expansion of f_C, written as $f_C(y_1, \ldots, y_q) = \sum_{S_1,\ldots,S_q \subseteq [\ell]} \widehat{f_C}(S_1, \ldots, S_q) \prod_{t=1}^q \prod_{j \in S_t} (y_t)_j$. Using the Fourier expansion of f_C, the Cauchy-Schwarz inequality, and Parseval's identity, we have

$$\begin{aligned}
\varepsilon^2 &\leq \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k}[b_i f_C(x_{v_1}, \ldots, x_{v_q})]^2 \\
&= \left(\sum_{S_1,\ldots,S_q \subseteq [\ell]} \widehat{f_C}(S_1, \ldots, S_q) \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k} \left[b_i \prod_{t=1}^q \prod_{j \in S_t} (x_{v_t})_j \right] \right)^2 \\
&\leq \left(\sum_{S_1,\ldots,S_q \subseteq [\ell]} \widehat{f_C}(S_1, \ldots, S_q)^2 \right) \left(\sum_{S_1,\ldots,S_q \subseteq [\ell]} \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k} \left[b_i \prod_{t=1}^q \prod_{j \in S_t} (x_{v_t})_j \right]^2 \right) \\
&= \left(\mathop{\mathbb{E}}_{y_1,\ldots y_q \leftarrow \{-1,1\}^\ell} [f_C(y_1, \ldots, y_q)^2] \right) \left(\sum_{S_1,\ldots,S_q \subseteq [\ell]} \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k} \left[b_i \prod_{t=1}^q \prod_{j \in S_t} (x_{v_t})_j \right]^2 \right) \\
&= \sum_{S_1,\ldots,S_q \subseteq [\ell]} \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k} \left[b_i \prod_{t=1}^q \prod_{j \in S_t} (x_{v_t})_j \right]^2 \\
&\leq 2^{q\ell} \max_{S_1,\ldots,S_q \subseteq [\ell]} \left\{ \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k} \left[b_i \prod_{t=1}^q \prod_{j \in S_t} (x_{v_t})_j \right]^2 \right\}
\end{aligned}$$

Thus we can find sets $R_1^C, \ldots, R_q^C \subseteq [\ell]$ and bit $t_C \in \{0,1\}$ such that

$$(-1)^{t_C} \mathop{\mathbb{E}}_{b \leftarrow \{-1,1\}^k} \left[b_i \prod_{t=1}^q \prod_{j \in S_t} (x_{v_t})_j \right] \geq \frac{\varepsilon}{2^{q\ell/2}} \geq \frac{\varepsilon}{2^{q-1}|\Sigma|^{q/2}} \ .$$

Reverting back from $\{-1,1\}$ to $\{0,1\}$ in the natural way, the last expression is equivalent to

$$\Pr_{b \leftarrow \{0,1\}^k} \left[t_C + \sum_{i=1}^q \langle \mathbf{1}_{R_1^C}, x_{v_i} \rangle = b_i \right] \geq \frac{1}{2} + \frac{\varepsilon}{2^q |\Sigma|^{q/2}} \ .$$

Thus we can form a new query set $C' \coloneqq \{(v_1, (\mathbf{1}_{R_1^C}, t_C)), (v_2, (\mathbf{1}_{R_2^C}, 0)), \ldots, (v_q, (\mathbf{1}_{R_q^C}, 0))\}$ for C' that recovers b_i with probability $1/2 + \varepsilon/(2^q|\Sigma|^{q/2})$. Indeed, this is how we construct our new hypergraphs $\mathcal{H}_1', \ldots, \mathcal{H}_k'$. Since we are mapping each query set to a new one, then we see that $|\mathcal{H}_i| = |\mathcal{H}_i'| \geq \varepsilon\delta n/q^2 \geq \varepsilon\delta n'/(4q^2|\Sigma|)$ for all $i \in [k]$. Furthermore, the query mapping preserves disjointness and size, implying that the new hypergraph is a collection of k q-uniform matchings. This finishes the proof. □

B Our Proof as a Black-box Reduction to 2-LDC Lower Bounds

In this appendix, we reinterpret our proof of Theorem 1 in the specific case of *linear* 3-LDCs by formulating it as a black-box reduction to existing linear 2-LDC lower bounds. Because we are reinterpreting the proof, we will assume familiarity with the proof in Sections 3 and 4. Formally, we show that our proof of Theorem 1 in fact provides the following transformation: given a *linear* 3-LDC C, we produce 2 different linear codes C_2 and C_3 corresponding to the 2-XOR instance g_b and 3-XOR instance f_b from Section 3, with the guarantee that at least one of these codes is a linear 2-LDC. We note that unlike Theorem 1, this reduction-based proof will only apply to *linear* 3-LDCs. However, in this case we will obtain slightly better dependencies on $\log n$, ε, and δ than that in Theorem 1; this comes entirely from the fact that 2-LDC lower bounds for linear codes have slightly better dependencies on ε and δ than 2-LDC lower bounds for general, nonlinear codes.

Our transformation naturally produces objects that are formally not quite linear 2-LDCs, which we call "weak LDCs", defined below.

Definition B.1 (Linear weak LDC)**.** Given a code $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$, we say that C is a linear (q,δ)-*weakly locally decodable code* (or, (q,δ)-wLDC) if C is a linear code and there are q-uniform hypergraph matchings $\mathcal{H}_1,\dots,\mathcal{H}_k$ over $[n]$ such that (1) $\sum_{i=1}^{k}|\mathcal{H}_i| \geq \delta n k$ for any $i \in [k]$, and (2) $C \in \mathcal{H}_i$, we have that $\bigoplus_{v\in C}\mathsf{C}(b)_v = b_i$ for all messages $b \in \{0,1\}^k$.

We note that we work with weak LDCs solely for notational convenience, as it is straightforward to observe that they are equivalent to LDCs, up to constant factors in parameters. Indeed, the difference between a weak LDC and a true LDC is that the weak LDC only requires that $\sum_{i=1}^{k}|\mathcal{H}_i| \geq \delta n k$, rather than the stronger condition that $|\mathcal{H}_i| \geq \delta n$ for all $i \in [k]$. So, by removing all hypergraphs $\mathcal{H}_i$ with $|\mathcal{H}_i| \leq \delta n/2$ and setting the corresponding b_i's to 0, we obtain a new code $\mathsf{C}'\colon \{0,1\}^{k'} \to \{0,1\}^n$ where $k' \geq \delta k$ and $|\mathcal{H}_i| \geq \delta n/2$ for all $i \in [k']$.

Regardless, we note that the linear 2-LDC lower bound of [GKST06], which here we will use as a black-box, holds for linear weak 2-LDCs as well.

Lemma B.2 (Lemma 3.3 of [GKST06])**.** *Any linear* $(2,\delta)$*-wLDC* $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ *satisfies* $n \geq 2^{\delta k}$.

As the main theorem in this section, we will prove the following theorem.

Theorem B.3. *Let* $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ *be a linear* $(3,\delta)$*-wLDC, and let* $d \in \mathbb{N}$*. Then, there are codes* $\mathsf{C}_2\colon \{0,1\}^{k_2} \to \{0,1\}^n$ *and* $\mathsf{C}_3\colon \{0,1\}^{k_3} \to \{0,1\}^N$ *such that either* C_2 *is a linear* $(2,\Omega(\delta\cdot\frac{d}{d+k}))$*-wLDC or* C_3 *is a linear* $(2,\Omega(\delta^2/d))$*-wLDC, where* $k_2,k_3 \geq k/2$*,* $N = \binom{2n}{\ell}$ *and* $\ell = \sqrt{n/k}/c$*, where* c *is an absolute constant.*

We note that by applying Lemma B.2 twice, we immediately obtain the following corollary.

Corollary B.4. *Let* $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ *be a* $(3,\delta)$*-linear LDC. Then,* $n \geq \Omega\left(\frac{\delta^6 k^3}{\log^4 k}\right)$.

Proof. Apply Theorem B.3 with $d = c\log_2 n/\delta$ for a sufficiently large constant c. If $k \leq d$, then we are done, so suppose that $k \geq d$. If C_2 is a linear weak $(2,\Omega(\delta\cdot\frac{d}{d+k}))$-LDC, then by Lemma B.2 we conclude that $\log_2 n \geq \Omega(\delta dk/(k+d)) \geq \Omega(\delta d)$, as $k+d \leq 2k$. As $d = c\log_2 n/\delta$ for a sufficiently large constant c, this is a contradiction.

It thus cannot be the case that C_2 is a linear weak $(2, \Omega(\delta \cdot \frac{d}{d+k}))$-LDC, and therefore it must be the case that C_3 is a linear weak $(2, \Omega(\delta^2/d))$-LDC. By Lemma B.2, this implies that $O(\sqrt{n/k}\log n) \geq \ell \log_2 n \geq \Omega(\delta^2/d \cdot k)$, and therefore we conclude that $n \geq \Omega(\delta^6 k^3/\log^4 n)$. Finally, we have $\log_2 n = \Theta(\log k)$ or else Corollary B.4 trivially holds, and so this finishes the proof. □

We now prove Theorem B.3.

Proof of Theorem B.3. Let $\mathsf{C}\colon \{0,1\}^k \to \{0,1\}^n$ be a linear $(3,\delta)$-wLDC, so that there exist 3-uniform hypergraph matchings $\mathcal{H}_1, \ldots, \mathcal{H}_k$ such that $\sum_{i=1}^k |\mathcal{H}_i| \geq \delta n k$, and for every $i \in [k]$ and $C \in \mathcal{H}_i$, it holds that $\bigoplus_{v \in C} \mathsf{C}(b)_v = b_i$ for all $b \in \{0,1\}^k$.

We now define the codes C_2 and C_3. Let $G_1, \ldots, G_k, \mathcal{H}'_1, \ldots, \mathcal{H}'_k$ denote the output of the hypergraph decomposition algorithm Lemma 3.2 applied with the parameter d chosen in the statement of Theorem B.3.

Constructing C_2. Let $L_2 \subseteq [k]$ be a subset of size $|L_2| \geq k/2$ to be specified later. We let $\mathsf{C}_2\colon \{0,1\}^{L_2} \to \{0,1\}^n$ be the code that encodes a message $b' \in \{0,1\}^{L_2}$ as $\mathsf{C}(b)$, where b is obtained by padding b' with 0's to obtain $b \in \{0,1\}^k$. Formally, $\mathsf{C}_2(b') \coloneqq \mathsf{C}(b)$, where $b \in \{0,1\}^k$ satisfies $b_i = b'_i$ for all $i \in L_2$ and $b_j = 0$ otherwise.

We will now show that if $\sum_{i=1}^k |G_i| \geq \delta n k/2$, then there exists a set $L_2 \subseteq [k]$ of size $|L_2| \geq k/2$ such that C_2 is a linear $(2, \Omega(\delta \cdot \frac{d}{d+k}))$-wLDC. Recall that each G_i is a bipartite matching on $[n] \times P$, where $P = \{p = (u,v) : \deg_{\mathcal{H}}(p) \geq d\}$, where $\mathcal{H} = \cup_{i=1}^k \mathcal{H}_i$. First, by duplicating elements of the set P, we can furthermore assume that each $p \in P$ appears not just in at least d edges across all G_i's, but also in at most $2d$ edges. Partition $[k]$ into $L_2 \cup R_2$, and without loss of generality assume $|L_2| \geq k/2$. For $i \in L_2$, let G'_i denote the graph on n vertices with edges $E_i = \{(u,v) : \exists p \in P, j \in R_2, (u,p) \in G_i, (v,p) \in G_j\}$. Observe that $\sum_{i \in L_2} |G'_i| \geq \Omega(\delta n k d)$ in expectation over a random partition $L_2 \cup R_2$, and hence there exists such a partition $L_2 \cup R_2$ with $\sum_{i \in L_2} |G'_i| \geq \Omega(\delta n k d)$.

Next, we observe that for any vertex $u \in [n]$ and $i \in L_2$, u has degree at most $2d + k$ in G'_i. Indeed, since the G_i's are matchings and each p appears in at most $2d$ edges, it follows that for each u, there are at most $2d$ edges (u,v) in G'_i formed from the edge (u,p) in G_i. Second, for each v, there are at most k edges (u,v) in G'_i, as these can only be formed from the edges (v,p) in G_j, for $j \in R_2$, and each G_j is matching so there is at most one edge per choice of $j \in R_2$. Hence, each G'_i has a matching M'_i of size at least $\Omega(|G'_i|/(d+k))$, and so $\sum_{i=1}^k |M'_i| \geq \Omega(\delta n k \cdot \frac{d}{d+k})$.

Finally, for each $i \in L_2$ and each edge $(u,v) \in M'_i$, it holds that $\mathsf{C}_2(b')_u \oplus \mathsf{C}_2(b')_v = b'_i$. Indeed, this is because $\mathsf{C}(b)$ satisfies $\mathsf{C}(b)_u \oplus \mathsf{C}(b)_p = b_i$ and $\mathsf{C}(b)_v \oplus \mathsf{C}(b)_p = b_j = 0$, where $p \in P$ is the shared pair used to add (u,v) to G'_i in the definition, $j \in R_2$, and $(u,p) \in G_i, (v,p) \in G_j$. We have thus shown that if $\sum_{i=1}^k |G_i| \geq \delta n k/2$, then C_2 is a linear $(2, \Omega(\delta \cdot \frac{d}{d+k}))$-wLDC.

Constructing C_3. Let $L_3 \subseteq [k]$ be a subset of size $|L_3| \geq k/2$ to be specified later. Let $\ell = \sqrt{n/k}/c$ for a sufficiently large constant c, and identify $N = \binom{2n}{\ell}$ with the collection of sets $\binom{[n]\times[2]}{\ell}$. We let $\mathsf{C}_3\colon \{0,1\}^{L_3} \to \{0,1\}^N$ be the code that encodes a message $b' \in \{0,1\}^{L_3}$ with the string $\mathsf{C}_3(b')$, where the S-th entry, for $S \in \binom{[n]\times[2]}{\ell}$, is

$$\mathsf{C}_3(b')_S \coloneqq \Big(\bigoplus_{u^{(1)} \in S} \mathsf{C}(b)_u\Big) \oplus \Big(\bigoplus_{v^{(2)} \in S} \mathsf{C}(b)_v\Big),$$

where $b \in \{0,1\}^k$ satisfies $b_i = b'_i$ for all $i \in L_3$ and $b_j = 0$ otherwise.

We now argue that if $\sum_{i=1}^k |\mathcal{H}'_i| \geq \delta nk/2$, then there exists a set $L_3 \subseteq [k]$ of size $|L_3| \geq k/2$ such that C_3 is a linear $(2, \Omega(\delta^2/d))$-wLDC. Recall that each $\mathcal{H}'_i$ is a 3-uniform hypergraph matching on n vertices, where $\deg_{\mathcal{H}'}(\{u,v\}) \leq d$ for all $u, v \in [n]$, where $\mathcal{H}' := \cup_{i=1}^k \mathcal{H}'_i$. Partition $[k]$ into $L_3 \cup R_3$, and without loss of generality assume $|L_3| \geq k/2$. Following Section 4, we set $\ell = \sqrt{n/k}/c$ for a sufficiently large constant c and let $B_i \in \mathbb{R}^{N \times N}$ for $i \in L_3$ be the matrices defined in Definition 4.4.

Let G''_i denote the graph with adjacency matrix B_i, i.e., for $S, T \in [N]$, we have (S,T) as an edge in G''_i if $B_i(S,T) \neq 0$. By Lemma 4.6, the max degree of any vertex in G''_i is at most $2d$. Hence, G''_i contains a matching M''_i where $|M''_i| \geq \Omega(|G''_i|/d)$. Now, since $|\mathcal{H}'| \geq \delta nk/2$, then by double counting, the number of clauses $C_1, C_2 \in \mathcal{H}'$ with $|C_1 \cap C_2| \geq 1$ is at least $\Omega(\delta^2 nk^2)$. Thus, by picking a random partition and using Lemma 4.7, we find that $\sum_{i=1}^k |G''_i| \geq \Omega(D\delta^2 nk^2)$ in expectation, where $D = 2\binom{2n-\ell}{\ell-4}$, and hence there is a partition $L_3 \cup R_3$ achieving this. By applying Fact 2.6, we see that $D/N \geq \Omega(\ell^2/n^2)$, and so we have $\sum_{i=1}^k |M''_i| \geq \Omega(\delta^2 Nk/d)$, using that $\ell = \sqrt{n/k}/c$.

It is now straightforward to observe that, for each $i \in L_3$ and $(S,T) \in M''_i$, it holds that $b'_i = \mathsf{C}_3(b')_S \oplus \mathsf{C}_3(b')_T$; indeed, this is because $\mathsf{C}_3(b')_S \oplus \mathsf{C}_3(b')_T = \mathsf{C}(b)_S \oplus \mathsf{C}(b)_T = b_i \oplus b_j = b'_i$, as $b'_i = b_i$ and $b_j = 0$ because $j \in R_2$. We have thus shown that if $\sum_{i=1}^k |\mathcal{H}'_i| \geq \delta nk/2$, then C_3 is a linear $(2, \Omega(\delta^2/d))$-wLDC.

By Lemma 3.2, we thus have that either $\sum_{i=1}^k |G_i| \geq \delta nk/2$ or $\sum_{i=1}^k |\mathcal{H}'_i| \geq \delta nk/2$. Hence, at least one of C_2 and C_3 must have the desired property, which finishes the proof. □

Remark B.5 (A note on the linearity of C). In Theorem B.3, we assumed that the code C was linear. The reason that this assumption is necessary is because of the following. The constraints used to locally decode C_2 and C_3 are obtained by XORing two clauses C_1 and C_2 in the original set of local constraints defining C. We then observe that by using $C_1 \oplus C_2$, we can decode, e.g., $b_i \oplus b_j$, and so by setting $\sim k/2$ of the b_j's to be hardcoded to 0, we have many constraints to recover b_i. The issue for nonlinear codes is that this "hardcoding" procedure does not work, as even though we can set b_j to be 0, the individual constraints C_1 and C_2 are only guaranteed to decode b_i and b_j, respectively, *in expectation* over a random choice of $b \in \{0,1\}^k$. Thus, when we hardcode some bits, we are no longer guaranteed that the derived constraint $C_1 \oplus C_2$ decodes b_i in expectation over the remaining "free" bits b_i for $i \in L$.

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