Electronic Colloquium on Computational Complexity, Report No. 187 (2024)

A $k^{\frac{q}{q-2}}$ Lower Bound for Odd Query Locally Decodable Codes from Bipartite Kikuchi Graphs

Oliver Janzer
oj224@cam.ac.uk
University of Cambridge

Peter Manohar
pmanohar@ias.edu
The Institute for Advanced Study

November 21, 2024

Abstract

A code $C\colon \{0,1\}^k \to \{0,1\}^n$ is a q-query 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 querying a corrupted string $\tilde{x}$ of the codeword $x = C(b)$ in at most q coordinates. For 2 queries, the Hadamard code is a 2-LDC of length $n = 2^k$, and this code is in fact essentially optimal [KW04, GKST06]. For $q \geq 3$, there is a large gap in our understanding: the best constructions achieve $n = \exp(k^{o(1)})$, while prior to the recent work of [AGKM23], the best lower bounds were $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ for q even and $n \geq \tilde{\Omega}(k^{\frac{q+1}{q-1}})$ for q odd.

The recent work of [AGKM23] used spectral methods to prove a lower bound of $n \geq \tilde{\Omega}(k^3)$ for $q = 3$, thus achieving the "$k^{\frac{q}{q-2}}$ bound" for an odd value of q. However, their proof does not extend to any odd $q \geq 5$. In this paper, we prove a q-LDC lower bound of $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ for *any* odd q. Our key technical idea is the use of an imbalanced bipartite Kikuchi graph, which gives a simpler method to analyze spectral refutations of *odd* arity XOR without using the standard "Cauchy–Schwarz trick" — a trick that typically produces random matrices with correlated entries and makes the analysis for odd arity XOR significantly more complicated than even arity XOR.

Keywords: Locally Decodable Codes, Spectral Refutation, Kikuchi Matrices

ISSN 1433-8092

Contents

1 Introduction

A (binary) locally decodable code (LDC) $C\colon \{0,1\}^k \to \{0,1\}^n$ is an error correcting code that admits a local decoding algorithm — for any message $b \in \{0,1\}^k$ and any string $\tilde{x} \in \{0,1\}^n$ obtained by corrupting the codeword $x = C(b)$ in a small constant fraction of coordinates, the local decoder is able to recover any bit b_i of b with good confidence while only reading a small number of coordinates of the corrupted codeword $\tilde{x}$. More formally, we say that C is (q, δ, ε)-locally decodable if the decoder only reads at most q bits of the corrupted string, and for any $\tilde{x}$ with Hamming distance $\Delta(x, \tilde{x}) := |\{u \in [n] : x_u \neq \tilde{x}_u\}| \leq \delta n$ and any input $i \in [k]$ to the decoder, the decoder recovers b_i with probability at least $\frac{1}{2} + \varepsilon$. Locally decodable codes were first formally defined in the work of [KT00], although they were instrumental components in the earlier proof of the PCP theorem [AS98, ALM+98], and have deep connections to complexity theory (see Section 7 in the survey of [Yek12]). Example applications include 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 question in coding theory is to determine the optimal blocklength n of a (q, δ, ε)-LDC as a function of k, the length of the message, and q, the number of queries, in the regime where q is constant and δ, ε are also constant. The work of [KT00] shows that there are no 1-query locally decodable codes unless k is constant, so the first nontrivial setting of q is $q = 2$. For 2-query locally decodable codes, we have an essentially complete understanding: the Hadamard code gives a 2-LDC with $n = 2^k$, and the works of [KW04, GKST06] show a lower bound of $n \geq 2^{\Omega(k)}$, which is therefore tight up to a constant in the exponent.

Unlike the case of $q = 2$, for $q \geq 3$ there is a large gap between the best-known upper and lower bounds on n. The best-known upper bound, i.e., construction, comes from matching vector codes [Yek08, Efr09], and achieves, in the case of $q = 3$, a blocklength of $n = \exp(\exp(O(\sqrt{\log k \log\log k})))$. This is $2^{k^{o(1)}}$, i.e., subexponential in k, which is substantially smaller than the Hadamard code, the code of optimal length for $q = 2$. More generally, for any constant $q = 2^r$, the works of [Yek08, Efr09] construct q-query locally decodable codes of length $n = \exp(\exp(O((\log k)^{1/r}(\log\log k)^{1-1/r})))$, which has a similar qualitative subexponential behavior.

On the other hand, the known lower bounds for $q \geq 3$ are substantially weaker. The original work of [KT00] proves that a q-LDC has blocklength $n \geq \Omega(k^{\frac{q}{q-1}})$. This was later improved by the work of [KW04], which showed that for even q, a q-LDC has blocklength $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$. For odd q, they observe that a q-LDC is also a $(q+1)$-LDC where $q+1$ is now even, and so their lower bound for even q trivially yields a bound of $n \geq \tilde{\Omega}(k^{\frac{q+1}{q-1}})$ for odd q.

The lower bounds of [KW04] remained, up to polylog(k) factors, the best lower bounds known until the recent work of [AGKM23], which used spectral methods developed in the work of [GKM22] for refuting constraint satisfaction problems to prove a lower bound of $n \geq \tilde{\Omega}(k^3)$ for 3-LDCs. This lower bound of [AGKM23] was the first improvement in any LDC lower bound by a poly(k) factor since the work of [KW04], and achieves the "$k^{\frac{q}{q-2}}$ bound" established for even q for the odd value of $q = 3$. However, the proof of [AGKM23] does not extend to any odd $q \geq 5$, and while

the spectral method approach of [AGKM23] was used in recent work of [KM24a] (and follow-ups [Yan24, AG24, KM24b]) to prove an exponential lower bound for 3-query locally *correctable* codes (LCCs) — a stronger variant of an LDC where the decoder must additionally be able to correct any bit x_u of the uncorrupted codeword — there have been no improvements in q-LDC lower bounds since [AGKM23]. In particular, because the proof of [AGKM23] does not extend to all odd $q \geq 5$, the best known lower bounds for q-LDCs are: (1) $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$, if q is even or $q = 3$, and (2) $n \geq \tilde{\Omega}(k^{\frac{q+1}{q-1}})$ if $q \geq 5$ is odd. Thus, a natural question to ask is: *can we prove a* $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ *lower bound for* q*-LDCs for all constant* q*?*

As the main result of this paper, we prove the following theorem, which establishes this lower bound.

Theorem 1. *Let* $C \colon \{0,1\}^k \to \{0,1\}^n$ *be a* (q, δ, ε)*-locally decodable code with* $q \geq 3$ *and* q *odd. Then,* $k \leq O_q(n^{1-\frac{2}{q}} \varepsilon^{-6-\frac{2}{q}} \delta^{-2-\frac{2}{q}} \log n)$*. In particular, if* q, δ, ε *are constants, then* $n \geq \Omega\left((k/\log k)^{\frac{q}{q-2}}\right)$.

The main contribution of Theorem 1 is that it improves the q-LDC lower bound for $q \geq 5$ from $n \geq \tilde{\Omega}(k^{\frac{q+1}{q-1}})$ to $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$, which is a $\mathrm{poly}(k)$ factor improvement. However, we additionally note that for $q = 3$, Theorem 1 has a better dependence on the lower order terms of $\log k, \delta, \varepsilon$ hidden in the $\tilde{\Omega}(\cdot)$ as compared to the result of [AGKM23], which showed the weaker bound of $n \geq \Omega\left(\frac{\varepsilon^{32}\delta^{16}k^3}{\log^6 k}\right)$.

As stated in [AGKM23], the proof techniques of [AGKM23] extend to any $q \geq 5$ *under the additional assumption that the code* C *satisfies some extra nice regularity properties.*[1] This condition arises for q odd but not q even for the following reason: in the proof, one defines a matrix where we would like to "evenly split" a degree-q monomial $x_{v_1} \dots x_{v_q}$ across the rows and columns of the matrix. When q even this is possible, as we can divide the monomial into two "halves". This property allows us to define a matrix with independent bits of randomness and obtain a somewhat simple proof. However, for q odd, the best possible split is of course $(\frac{q-1}{2}, \frac{q+1}{2})$, which is slightly imbalanced. To handle this issue of imbalance, [AGKM23] uses the standard "Cauchy–Schwarz trick" developed in the context of spectral refutation algorithms for constraint satisfaction problems precisely to tackle this issue of imbalance. The "Cauchy–Schwarz trick" produces degree $2(q-1)$ monomials, which are even, but at the cost of making the randomness *dependent*. This dependence in the randomness makes the analysis for odd q considerably more technical than the more straightforward analysis for even q, and is where the aforementioned additional assumption on the code is needed. The fact that even q is substantially more easier to handle from a technical perspective compared to odd q is a reoccurring theme in the CSP refutation literature that has appeared in many prior works [CGL07, AOW15, BM16, RRS17, AGK21, GKM22, KM24a].

Our main technical contribution is the introduction of an *imbalanced* matrix, or equivalently a *bipartite* graph for odd arity instances, that allows us to refute certain odd arity instances *without* using the Cauchy–Schwarz trick. By using an imbalanced matrix and bypassing the Cauchy–Schwarz trick,

[1] In fact, as we mention in Section 2.2, it turns out that the proof strategy of [AGKM23] extends easily to the case of $q = 5$ without the need for any additional assumption, contrary to what is claimed in [AGKM23]. The fact that this was missed by [AGKM23] appears to be an oversight on their part. Nonetheless, their approach does break down for $q \geq 7$.

we maintain independence in the randomness in our matrix (as opposed to introducing correlations), which makes the proof considerably simpler. The simpler proof has the additional advantage that, as mentioned earlier, in the case of $q = 3$ we can improve on the lower bound of [AGKM23] by a $\log k \cdot \mathrm{poly}(1/\varepsilon, 1/\delta)$ factor. Our use of a bipartite graph in the proof is perhaps surprising, as it is contrary to the conventional wisdom that symmetric matrices (i.e., normal, non-bipartite graphs) ought to produce the best spectral certificates. Indeed, the purpose of "Cauchy–Schwarz trick" is to turn an odd arity instance into an even arity instance so that we can represent it with a balanced matrix, with the expectation (that is true in many cases) that the balanced matrix will produce a better spectral certificate.

The bipartite graph that we produce is a Kikuchi graph, i.e., a carefully chosen induced subgraph of a Cayley graph on the hypercube. As we note in Remark 2.5, bipartite Kikuchi graphs have appeared in prior works, namely [Yan24, KM24b]. However, the graphs in those works can be converted to non-bipartite graphs via a straightforward application of the Cauchy–Schwarz inequality, and so they are not "inherently bipartite". To our knowledge, our work is the first work to use such a graph that is *inherently* bipartite, meaning that no easy conversion to a non-bipartite graph via the Cauchy–Schwarz inequality appears to exist.

Concurrent work. In concurrent work, [BHKL24] also proves a $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ lower bound for q-query locally decodable codes for odd q. Their bound is slightly weaker compared to Theorem 1, as it has a worse $\log n$ dependence. Namely, for constant ε, δ, [BHKL24] proves that $k \leq O(n^{1-2/q}(\log n)^4)$ for nonlinear codes and $k \leq O(n^{1-2/q}(\log n)^2)$ for linear codes, whereas Theorem 1 proves that $k \leq O(n^{1-2/q} \log n)$ for both nonlinear and linear codes, which is a stronger bound in both cases.

2 Proof Overview

In this section, we give an overview of our proof and the techniques that we use. We will start with a thorough summary of the approach of [AGKM23], first for the (easier) case of even q, and then for the more involved case of $q = 3$. Then, we will explain our approach using bipartite Kikuchi graphs.

For the purpose of this overview, we will assume for simplicity that the code C is linear, although we note that the proof for nonlinear codes does not change in any meaningful way.

2.1 The approach of [AGKM23] for even q

By standard reductions (Fact 3.5), for any linear q-LDC $C\colon \{0,1\}^k \to \{0,1\}^n$, there exist q-uniform hypergraph matchings (Definition 3.1) $H_1, \ldots, H_k$ on the vertex set $[n]$, each with $|H_i| = \delta n$ hyperedges, such that for each $i \in [k]$ and each hyperedge $C \in H_i$, it holds that $\sum_{v \in C} x_v = b_i$ when $x = C(b)$. One should think of the hyperedges H_i as the set of query sets that the decoder may query on input i. That is, the decoder, when given input $i \in [k]$, simply chooses a random $C \leftarrow H_i$, reads $x|_C$, and then outputs $\sum_{v \in C} x_v$. The linear constraints $\sum_{v \in C} x_v = b_i$ that are satisfied by $x = C(b)$ for all $b \in \{-1,1\}^k$ imply that the decoder succeeds in correctly recovering b_i with probability 1 on an uncorrupted codeword.

Switching from $\{0,1\}$-notation to $\{-1,1\}$-notation via the map $0 \mapsto 1$ and $1 \mapsto -1$, the above implies that for any $b \in \{-1,1\}^k$, the system of constraints given by $\prod_{v \in C} x_v = b_i$ for each $i \in [k]$ and $C \in H_i$ is satisfiable, with $x = C(b)$ being a satisfying assignment. This implies that the degree-q polynomial $\Phi_b(x) \coloneqq \sum_{i=1}^k b_i \sum_{C \in H_i} \prod_{v \in C} x_v$ has value $\mathrm{val}(\Phi_b) \coloneqq \max_{x \in \{-1,1\}^n} \Phi_b(x) = \sum_{i=1}^k |H_i| = \delta n k$ for all $b \in \{-1,1\}^k$. Indeed, by setting $x = C(b)$, we have that $\prod_{v \in C} x_v = b_i$ for each $i \in [k]$ and $C \in H_i$, and so $\Phi_b(C(b)) = \sum_{i=1}^k |H_i| = \delta n k$. Thus, to prove a lower bound on n, it suffices to show that for any $H_1, \dots, H_k$ of size δn, if n is too small, then there exists $b \in \{-1,1\}^k$ such that $\mathrm{val}(\Phi_b) < \delta n k$.

We do this by bounding $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\Phi_b)]$ using a *spectral certificate*. The certificate is as follows. First, we define the Kikuchi matrix/graph A_C, as follows.

Definition 2.1 (Basic Kikuchi matrix/graph for q even)**.** Let $\ell \geq q$ be a positive integer (which we will set to $n^{1-2/q}$ eventually), and let $C \in \binom{[n]}{q}$. Let A_C be the matrix with rows and columns indexed by sets $S \in \binom{[n]}{\ell}$ where $A_C(S,T) = 1$ if $S \oplus T = C$ and $A_C(S,T) = 0$ otherwise. Here, $S \oplus T$ denotes the symmetric difference of S and T, which is $\{u : (u \in S \wedge u \notin T) \vee (u \notin S \wedge u \in T)\}$. We will at times refer to the matrix A_C as a graph (where we identify A_C with the graph with adjacency matrix A_C), and then we will refer to the nonzero entries (S,T) as edges.

Notice that the condition that $S \oplus T = C$ is equivalent to the existence of a partition $C = C_1 \cup C_2$ into two sets of size $\frac{q}{2}$ such that $S \cap C = C_1$, $T \cap C = C_2$, and $S \setminus C_1 = T \setminus C_2$. Namely, $S \oplus T = C$ if and only if we can split the hyperedge C *evenly* across S and T — notice that here we crucially require that q is even for the matrix A_C to have a single nonzero entry!

The matrix A_C is a Kikuchi matrix, first introduced in the work of [WAM19] for the problem of tensor PCA, and has the following nice properties: (1) the matrix A_C has exactly $D = \binom{q}{q/2}\binom{n-q}{\ell-q/2}$ nonzero entries, and (2) for each $x \in \{-1,1\}^n$, letting $z \in \{-1,1\}^{\binom{n}{\ell}}$ denote the vector where $z_S \coloneqq \prod_{v \in S} x_v$, we have $z^\top A_C z = D \prod_{v \in C} x_v$. These two properties allow us to use the A_C's as a "basis" to express any homogeneous degree-q polynomial in variables $x \in \{-1,1\}^n$ as a quadratic form on a linear combination of A_C's. Namely, if we let $A_i \coloneqq \sum_{C \in H_i} A_C$ and $A = A_b \coloneqq \sum_{i=1}^k b_i A_i$ (mimicking the definition of Φ_b), then we have $z^\top A z = D\Phi_b(x)$ for every $x \in \{-1,1\}^n$, where z is defined as before. We can thus express $\Phi_b(x)$ as a quadratic form on the matrix A, and so we have shown that $\delta n k \leq \mathrm{val}(\Phi_b) \leq \|A\|_2 \cdot \binom{n}{\ell}$. So, to finish the proof, it remains to bound $\mathbb{E}_b[\|A\|_2]$.

As each b_i is chosen independently from $\{-1,1\}$, the matrix $A = \sum_{i=1}^k b_i A_i$ is a Matrix Rademacher series, and so we can bound its spectral norm using the Matrix Khintchine inequality (Fact 3.6). This implies that $\mathbb{E}_b[\|A\|_2] \leq O(\Delta\sqrt{k\ell \log n})$, where Δ is the maximum number of 1's in a row any of the A_i's. As the A_i's are symmetric matrices with entries in $\{0,1\}$, we can view them as adjacency matrices of graphs. With this perspective, Δ is simply the maximum degree of a vertex S in any of the A_i's.

The maximum degree Δ can never be smaller than the average degree in an A_i, which is $\delta n D / \binom{n}{\ell}$. Thus, if each A_i is approximately regular, so that the maximum degree is on the same order of magnitude as the average degree, then we would be able to conclude that

$$\delta n k D \leq \binom{n}{\ell} \mathbb{E}_b[\|A\|_2] \leq \binom{n}{\ell} \cdot \frac{\delta n D}{\binom{n}{\ell}} \cdot O(\sqrt{k\ell \log n})$$

$$\implies k \le O(\ell \log n)\,.$$

Ideally, we would like to take ℓ to be as small as possible to now get the best possible bound on k. However, the maximum degree Δ is always at least 1, and so for Δ to be on the same order of magnitude as the average degree, we must have average degree $\ge \Omega(1)$. The average degree is $\delta n D/\binom{n}{\ell}$, which a simple calculation shows is roughly $\delta n \left(\frac{\ell}{n}\right)^{q/2}$, and so we need to take $\ell \ge n^{1-2/q}$. This means that our potential bound is $k \le O(n^{1-2/q}\log n)$, i.e., $n \ge \tilde{\Omega}(k^{\frac{q}{q-2}})$, as desired.

Finding an approximately regular subgraph: row pruning. We have shown that if each graph A_i is approximately regular, meaning that its maximum degree is on the same order of magnitude as its average degree, then we can prove $n \ge \tilde{\Omega}(k^{\frac{q}{q-2}})$. Unfortunately, the graph A_i is not approximately regular, *even though the underlying hypergraph H_i is a matching*, i.e., H_i is as "regular" as possible. A naive way to try to enforce this "approximately regular" property is to simply remove all vertices S with large degree in A_i (along with their adjacent edges), producing a new graph B_i with max degree bounded by $O(1)$ times the average degree of A_i. However, for a general graph, this deletion process may delete most (or all!) of the edges, resulting in a considerable drop in the average degree. So, the resulting graph B_i need not be approximately regular. Crucially, *because H_i is a matching*, we can show that this process in fact only deletes a $o(1)$-fraction of the edges, and so the average degree is essentially unchanged. This means that the graph B_i *is* indeed approximately regular, and so we can use the B_i's in place of the A_i's to finish the proof.

The above vertex/edge deletion step is typically called the "row pruning" step, so-named because it prunes rows (and columns) of the matrix A_i, and has appeared in many prior works that analyze spectral norms of Kikuchi matrices. While at first glance this step may appear to be a mere technical annoyance, it is in fact the most critical part of the entire proof. In this case of the above proof, we note that this is the only step that uses that the H_i's are matchings, and if the H_i's are not matchings then the lower bound is clearly false. In fact, in the entire proof above, one should view all the steps up until the row pruning step as *generic* and dictated by the polynomial Φ_b whose value we wish to bound, and the row pruning step is the key part of the proof that determines if the approach succeeds in obtaining a strong enough bound on $\mathbb{E}_b[\mathrm{val}(\Phi_b)]$.

As observed in [AGKM23], one can also view the above proof as giving a reduction from a q-LDC to a 2-LDC for even q. In this viewpoint, the row pruning step is the crucial part of the proof that shows that the object produced by the reduction is in fact a 2-LDC.

2.2 The approach of [AGKM23] for $q = 3$ and why it fails for odd $q \ge 7$

We now recall the approach of [AGKM23] for $q = 3$ and explain why its natural generalization to odd $q \ge 7$ fails. As briefly mentioned earlier, the reason the previous proof does not succeed for odd q is because the matrix A_C has no nonzero entries when $|C|$ is odd. This is because the row sets S and the column sets T have exactly the same size ℓ; the Kikuchi graph would have nonzero entries if we, e.g., simply allowed $|T| = \ell + 1$. Namely, we can make the following definition.

Definition 2.2 (Naive imbalanced Kikuchi matrix for odd q)**.** Let q be odd and let $C \in \binom{[n]}{q}$. Let $\ell \ge q$ be a positive integer. Let A'_C be the matrix with rows indexed by sets $S \in \binom{[n]}{\ell}$ and columns

indexed by sets $T \in \binom{[n]}{\ell+1}$, where $A'_C(S,T) = 1$ if $S \oplus T = C$ and otherwise $A'_C(S,T) = 0$.

Analogously to Definition 2.1, if the (S,T)-th entry of A'_C is nonzero, then $|S \cap C| = \frac{q-1}{2}$ and $|T \cap C| = \frac{q+1}{2}$. This imbalance causes the average left degree of $A'_i := \sum_{C \in H_i} A'_C$ to be roughly $\delta n \left(\frac{\ell}{n}\right)^{(q-1)/2}$, while the average right degree is $\delta n \left(\frac{\ell}{n}\right)^{(q+1)/2}$. In order for the row pruning step to have any hope for success, we need both of these quantities to be at least 1, which requires taking $\ell \geq n^{1-\frac{2}{q+1}}$. In fact, using this asymmetric matrix precisely reproduces the $n \geq \tilde{\Omega}(k^{\frac{q+1}{q-1}})$ bound.

The first key step in the proof of [AGKM23] for $q = 3$ is to use the "Cauchy–Schwarz trick" from the CSP refutation literature: we construct a new system of constraints by first taking two constraints $x_u x_{v_1} x_{v_2} = b_i$ and $x_u x_{w_1} x_{w_2} = b_j$ *that both contain the same variable* x_u and then we multiply them together to derive a new constraint $x_{v_1} x_{v_2} x_{w_1} x_{w_2} = b_i b_j$, using that $x_u^2 = 1$ since $x_u \in \{-1, 1\}$. Crucially, the arity of the derived monomial is 4 (or more generally $2(q-1)$),[2] which is now even. We thus define a new polynomial Ψ_b of even degree for the derived instance:

$$\Psi_b(x) := \sum_{i \neq j} b_i b_j \sum_{u \in [n]} \sum_{\substack{(u,v_1,v_2) \in H_i \\ (u,w_1,w_2) \in H_j}} b_i b_j x_{v_1} x_{v_2} x_{w_1} x_{w_2} \,.$$

A simple application of the Cauchy–Schwarz inequality relates $\mathrm{val}(\Phi_b)$ and $\mathrm{val}(\Psi_b)$, and hence this derivation process is typically called the "Cauchy–Schwarz trick". The main drawback is that in the derived constraints, the "right-hand sides" are products $b_i b_j$, and we have introduced correlations in the right-hand sides.

We can now use the Kikuchi graphs A_C (Definition 2.1) for each *derived constraint* C, as the derived constraints have even arity. However, we will make one small, but crucial change. For a derived constraint $x_{v_1} x_{v_2} x_{w_1} x_{w_2}$, where v_1, v_2 "come from" one hyperedge (u, v_1, v_2) and w_1, w_2 "come from" the other hyperedge (u, w_1, w_2), we view this constraint as two pairs $(\{v_1, v_2\}, \{w_1, w_2\})$, and for an edge (S,T) in the graph $A_{(\{v_1,v_2\},\{w_1,w_2\})}$, we require that S contains one element from each of $\{v_1, v_2\}$ and $\{w_1, w_2\}$, and that T contains the other element from each. That is to say, we *evenly split* the variables from the underlying (original) hyperedges across the row set S and column set T. The fact that we split elements evenly is crucial for the row pruning step that we will discuss shortly.

With the above definition of the matrix $A_{(\{v_1,v_2\},\{w_1,w_2\})}$, we can then make the following definitions. First, we partition $[k]$ randomly into two sets $L \cup R$, with $|L| \geq \frac{k}{2}$ without loss of generality. Then, we let $A_{i,j} := \sum_{u \in [n]} \sum_{\substack{(u,v_1,v_2) \in H_i \\ (u,w_1,w_2) \in H_j}} A_{(\{v_1,v_2\},\{w_1,w_2\})}$, $A_i := \sum_{j \in R} b_j A_{i,j}$ and $A = \sum_{i \in L} b_i A_i$. The random partition of $[k]$ into $L \cup R$ is a nice trick used in [AGKM23] that makes the matrix A be the sum of mean 0 independent random matrices. At this point, we can now take $\ell = n^{1-2/q} = n^{1/3}$ and apply similar steps as done in the case of q even to finish the proof, provided that the "approximately regular" condition can be made to hold for each graph A_i, i.e., the row pruning step succeeds.

[2]The degree may be smaller if the constraints share at least 2 variables, but this would reduce the degree further and so it is only "better" for us. There are several simple ways to handle this issue, but we will ignore this technicality for the purpose of simplifying this proof overview.

Finding an approximately regular subgraph: row pruning. Let us now discuss the row pruning step for the matrices A_i. Unlike in the even case, the constraint hypergraph that defines the matrix A_i is no longer a matching. Instead, the edges in the graph A_i "come from" tuples $(u, \{v_1, v_2\}, \{w_1, w_2\})$ where $(u, v_1, v_2) \in H_i$ and $(u, w_1, w_2) \in \cup_{j \in R} H_j$ — here, u is the shared variable that is "canceled" by multiplying the two constraints together. To find an approximately regular subgraph of A_i, intuitively we need to show that a typical vertex has degree roughly equal to the average degree. We can try to understand how concentrated the degrees are in A_i by computing the variance of $\deg_i(S)$, the degree of S in A_i, when S is chosen uniformly at random (see Lemma 5.6 for a formal calculation that is closely related). Here, it is crucial that we have split the uncanceled variables $\{v_1, v_2\}$ of the hyperedge (u, v_1, v_2) evenly across S and T because if we had not, then any set S that contains both $\{v_1, v_2\}$ should[3] have degree $\Omega(k)$. This is much larger than the average degree, which one can show is $n^{-1/3}k$, and happens with probability $n^{-1/3}$: high enough to dominate the variance.

In fact, even when we use the even split, $\mathrm{Var}(\deg_i(S))$ may still be too large. However, from the calculation of $\mathrm{Var}(\deg_i(S))$, we can extract the following natural combinatorial condition that, if satisfied, will make the variance small enough to finish the proof: we require that each *pair* of variables $\{u, w\}$ appears in at most $d_2 := (\ell/n)^{\frac{1}{2}}k = n^{-1/3}k$ hyperedges in $\cup_{j \in R} H_j$. However, the hypergraph $\cup_{j \in R} H_j$ is a union of matchings — it is not a matching itself — and so it is quite possible that there are pairs of variables $\{u, w\}$ that appear in, say, $\Omega(k)$ hyperedges in $\cup_{j \in R} H_j$.[4] In fact, if many such "heavy pairs" $\{u, w\}$ exist, then we are unable to show that the graph A_i has an approximately regular subgraph, and the above proof fails!

Nonetheless, the above proof still accomplishes something nontrivial. For $q = 3$, we obtain a proof that $k \le \tilde{O}(n^{1/3})$ under the additional assumption that each pair $\{u, w\}$ of variables appears in at most $n^{-1/3}k$ hyperedges in $\cup_{j \in [k]} H_j$.[5] More generally, for larger odd q, we can show that $k \le \tilde{O}(n^{1-2/q})$ under the additional assumption that for any set Q of size $|Q| = s$ where $2 \le s \le \frac{q+1}{2}$, the set Q appears in at most $d_s := (\ell/n)^{s-\frac{3}{2}}k = n^{-\frac{2s}{q}+\frac{3}{q}}k$ hyperedges in $\cup_{j \in [k]} H_j$.

Removing the heavy pairs assumption. The final step in the proof of [AGKM23] is to remove this assumption by using the *hypergraph decomposition* method of [GKM22]. For each heavy pair $\{u, w\}$, we create a new "big variable" p and replace all hyperedges (u, w, v) with a new hyperedge (p, v). Then, we create a new set of derived constraints by canceling the heavy pair variables p, resulting in a new degree-2 polynomial whose value we can then bound.[6] So, if there are many heavy pairs, then we can produce a degree-2 polynomial, and otherwise we already win via the degree-4 polynomial.

The hypergraph decomposition strategy fails for $q \ge 7$. The above approach to handling heavy

[3] Formally, it has degree at least the number of hyperedges $(u, w_1, w_2) \in \cup_{j \in R} H_j$ that contain the variable u, and this is typically $\Omega(k)$. For example, it is $\Omega(k)$ for every u if the H_i's are random hypergraph matchings.

[4] Because the H_j's are matchings and $|R| \le k$, even a single variable u cannot appear in more than k hyperedges. This is why we do not encounter a "heavy singleton" condition.

[5] As we do not know R in advance, we must impose a global condition on $\cup_{j \in [k]} H_j$ instead of $\cup_{j \in R} H_j$. However, as we expect $|R|$ to be about $k/2$, this is also only off by a constant factor.

[6] Formally, the proof of [AGKM23] proceeds slightly differently and uses a bipartite graph, although it is equivalent to this.

pairs suggests a natural strategy to handle larger heavy sets. Namely, let $d_s := n^{-\frac{2s}{q}+\frac{3}{q}}k$ be the "threshold for heavy sets Q of size s" that we found via the variance calculation. For each $2 \le s \le \frac{q+1}{2}$, we let P_s denote the set of heavy Q's of size s. Then, for each heavy set Q, we can introduce a new variable p and replace all hyperedges $C \in \cup_{j\in[k]} H_j$ containing Q where Q is the largest heavy set (ties broken arbitrarily) with the hyperedge $(p, C \setminus Q)$. This will produce, for each $i \in [k]$, a hypergraph $H_i^{(s)}$ where each hyperedge in $H_i^{(s)}$ has the form (p, C') where $|C'| = q - s$ and $p \in P_s$ is a new variable.

The derivation strategy of the "Cauchy–Schwarz trick" now suggests that we should, for each s, group the hypergraphs $H_1^{(s)}, \dots, H_k^{(s)}$ together and derive constraints by canceling the new variables p. Namely, we take two hyperedges (p, C) and (p, C') that use the same p and combine them to produce the derived constraint (C, C') that has arity $2(q - s)$. Once again, we can define an analogous Kikuchi matrix and attempt to complete the proof, and the success of this strategy is determined by whether or not the row pruning step goes through.

It turns out that, for $q = 5$, this simple generalization does indeed succeed. We suspect that this was perhaps missed by [AGKM23] because the thresholds d_s are rather delicate, and the proof breaks if we set, e.g., $d_2 = n^{-1/5} \cdot n^{1-\frac{2}{5}} \log n$ as opposed to $n^{-1/5}k$ (recall that we expect k to be about $n^{1-\frac{2}{5}} \log n$ as this is the lower bound that we are shooting for).

However, for $q \ge 7$, this proof strategy fails. The first case that breaks is for $q = 7$ and $s = 4$, i.e., we have produced derived constraints (C, C') of arity 6 by canceling a heavy 4-tuple. The issue is that the hypergraphs $H_1^{(4)}, \dots, H_k^{(4)}$, which have hyperedges containing $q - s = 7 - 4 = 3$ original variables from $[n]$, may still contain heavy pairs or triples. It turns out that, for any odd q, the cases of $s = 2$ and $s = 3$ are always fine, so this problem does not arise for $q = 5$ (recall that $2 \le s \le \frac{q+1}{2}$, which is 3 when $q = 5$).

2.3 Recursive hypergraph decomposition and Kikuchi graphs for partite XOR

While the strategy described above fails for $q \ge 7$, there is again a natural next step to try. The problem with, e.g., the case of $q = 7$ and $s = 4$, is that the hypergraphs $H_1^{(4)}, \dots, H_k^{(4)}$ may still contain heavy pairs or triples. So, we can simply recurse and decompose these hypergraphs again to produce hypergraphs $H_1^{(4,3)}, \dots, H_k^{(4,3)}$ and $H_1^{(4,2)}, \dots, H_k^{(4,2)}$. Here, hyperedges in $H_i^{(4,3)}$ have the form $(p^{(4)}, p^{(3)})$ where $p^{(4)} \in P_4$ is a heavy 4-tuple and $p^{(3)} \in P_3$ is a heavy triple, and hyperedges in $H_1^{(4,2)}, \dots, H_k^{(4,2)}$ have the form $(p^{(4)}, p^{(2)}, v)$, where $p^{(4)} \in P_4$, $p^{(2)} \in P_2$, and $v \in [n]$. (Because $\cup_{i=1}^k H_i$ is the union of matchings, each variable v appears in at most k hyperedges, so it is not possible to have a heavy singleton.)

For this overview, let us consider the case of $H_1^{(4,2)}, \dots, H_k^{(4,2)}$. We need to bound the value of $\Phi_b^{(4,2)}(x, y)$, defined as

$$\Phi_b^{(4,2)}(x, y) := \sum_{i=1}^{k} b_i \sum_{(p^{(4)},p^{(2)},v)\in H_i^{(4,2)}} y_{p^{(4)}} y_{p^{(2)}} x_v \,.$$

As before, we can now derive constraints using the Cauchy–Schwarz trick. Namely, we can take two hyperedges $(p^{(4)}, p_1^{(2)}, v_1) \in H_i^{(4,2)}$ and $(p^{(4)}, p_2^{(2)}, v_2) \in H_j^{(4,2)}$ that share the same heavy 4-tuple

$p^{(4)}$, and then form the derived hyperedge $((p_1^{(2)}, v_1), (p_2^{(2)}, v_2))$.

Now, the approach of [AGKM23] breaks down. To use their Kikuchi graph, we need to be able to derive constraints that only use the original variables $[n]$. But, the above derived constraints still use "heavy pair" variables. One could try to, e.g., combine the derived constraint $((p_1^{(2)}, v_1), (p_2^{(2)}, v_2))$ with some constraint in $H_{j'}^{(4,2)}$ that also contains the heavy pair $p_1^{(2)}$, but such a constraint will be of the form $(p_2^{(4)}, p_2^{(2)}, v_3)$, i.e., it will have a new heavy 4-tuple. So, the new derived constraint will be $((p_1^{(2)}, v_1), v_2, (p_2^{(4)}, v_3))$, and we have the same problem again. Furthermore, we cannot try to combine different "types" of hypergraphs, e.g., combine constraints in $H_j^{(2)}$ for some j with constraints in $H_i^{(4,2)}$ for some i, as it could be the case that after the hypergraph decomposition step, *most* (or all) of the original δnk hyperedges are placed in, e.g., $H_i^{(4,2)}$ for some i, and so all hypergraphs of a different "type" are empty.

Let us now explain our approach to handle this problem. We need to design a Kikuchi matrix for hyperedges that are *partite*: each hyperedge in $\cup_{i=1}^{k} H_i^{(4,2)}$ has two vertices from the vertex set P_2 and 2 vertices from the vertex set $[n]$. We introduce the following Kikuchi graph in this work for partite hypergraphs, which is defined as follows. For a derived constraint $((p_1^{(2)}, v_1), (p_2^{(2)}, v_2))$, we let the matrix $A_{((p_1^{(2)}, v_1), (p_2^{(2)}, v_2))}$ be the matrix indexed by *pairs* of sets S_1 and S_2, where $S_1 \subseteq [n]$ has size ℓ and $S_2 \subseteq P_2$ also has size ℓ. That is, each row has a set for each distinct set of variables, which are $[n]$ and P_2. By analogy to the earlier definition of A_C, we should set $A_{((p_1^{(2)}, v_1), (p_2^{(2)}, v_2))}((S_1, S_2), (T_1, T_2)) = 1$ if $S_1 \oplus T_1 = \{v, v'\}$ and $S_2 \oplus T_2 = \{p_1^{(2)}, p_2^{(2)}\}$, and furthermore we require that the variables "coming from" the underlying original hyperedges in $H^{(4,2)}$ are split *as evenly as possible* across the rows and columns. Namely, we require that either $v_1 \in S_1, p_2^{(2)} \in S_2$ and $v_2 \in T_1, p_1^{(2)} \in T_2$, or vice versa, so that the variables $(v_1, p_1^{(2)})$ "coming from" the first underlying hyperedge are split across the row and column, and likewise for the second underlying hyperedge.

Analogously to the definitions in Section 2.2, we can randomly partition $[k]$ into $L \cup R$ and let $A_{i,j} := \sum_{p^{(4)} \in P_4} \sum_{\substack{(p^{(4)}, p_1^{(2)}, v_1) \in H_i^{(4,2)} \\ (p^{(4)}, p_2^{(2)}, v_2) \in H_j^{(4,2)}}} A_{((p_1^{(2)}, v_1), (p_2^{(2)}, v_2))}$, $A_i := \sum_{j \in R} b_j A_{i,j}$ and $A = \sum_{i \in L} b_i A_i$. A straightforward calculation shows that $\mathbb{E}_b[\|A\|_2]$ provides an upper bound on $\mathrm{val}(\Phi_b^{(4,2)})$. Thus, to determine whether or not this matrix A is good enough to prove the desired $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ bound, we need to argue that the row pruning step holds, i.e., each A_i can be made approximately regular.

It turns out that, while the calculations are substantially more complicated that those appearing in [AGKM23], this approach does in fact work, provided that we adjust the thresholds d_s slightly. We need to set $d_s = n^{-\frac{2s}{q}+\frac{2}{q}}k$ (instead of $n^{-\frac{2s}{q}+\frac{3}{q}}k$) for $2 \leq s \leq \frac{q-1}{2}$, while keeping $d_{\frac{q+1}{2}} = n^{\frac{2}{q}-1}k$ the same. This allows us to prove the $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ bound for $q = 7$.

Generalizing our $q = 7$ approach to all odd q. We can now generalize our above approach to all odd q as follows. The output of the recursive hypergraph decomposition step yields hypergraphs $H_1^{(s,t_1,\ldots,t_r)}, \ldots, H_k^{(s,t_1,\ldots,t_r)}$, where $2 \leq s \leq \frac{q+1}{2}$ is an integer, $s \geq t_1 \geq \cdots \geq t_r \geq 2$ are all positive integers, and $t_1 + t_2 + \cdots + t_r + s \leq q$. This notation means that each hypergraph $H_i^{(s,t_1,\ldots,t_r)}$ contains hyperedges of the form $(p^{(s)}, p^{(t_1)}, \ldots, p^{(t_r)}, C)$, where each $p^{(t_z)} \in P_{t_z}$, i.e., it is a heavy t_z-tuple,

$p^{(s)} \in P_s$ is a heavy s-tuple, and $C \subseteq [n]$ has size $q - s - t_1 - \cdots - t_r$ and is the set of remaining "original variables" from $[n]$. We then apply the Cauchy–Schwarz trick to form derived constraints, which now take the form $((p_1^{(t_1)}, \ldots, p_1^{(t_r)}, C_1), (p_2^{(t_1)}, \ldots, p_2^{(t_r)}, C_2))$. We define the Kikuchi matrix $A_{((p_1^{(t_1)},\ldots,p_1^{(t_r)},C_1),(p_2^{(t_1)},\ldots,p_2^{(t_r)},C_2))}$ analogously to the case of $H^{(4,2)}$ shown above for $q = 7$. Namely, we make the following definition.

Definition 2.3 (Kikuchi matrices for partite hypergraphs)**.** Let $((p_1^{(t_1)}, \ldots, p_1^{(t_r)}, C_1), (p_2^{(t_1)}, \ldots, p_2^{(t_r)}, C_2))$ be a derived constraint where $p_1^{(t_z)}, p_2^{(t_z)} \in P_{t_z}$ for each $z \in [r]$ and $C_1, C_2 \subseteq [n]$ have size $q - s - \sum_{z=1}^{r} t_z$. Let $A_{((p_1^{(t_1)},\ldots,p_1^{(t_r)},C_1),(p_2^{(t_1)},\ldots,p_2^{(t_r)},C_2))}$ be the matrix indexed by tuples of sets $(S_1, \ldots, S_r, S)$ where for $z \in [r]$, $S_z \subseteq P_z$ has size ℓ and $S \subseteq [n]$ has size ℓ. The matrix has a 1 in the $((S_1, \ldots, S_r, S), (T_1, \ldots, T_r, T))$-th entry if for each, $S_z \oplus T_z = \{p_1^{(t_z)}, p_2^{(t_z)}\}$, and $S \oplus T = C_1 \oplus C_2$. Otherwise, the entry is 0.

In order for the row pruning step to go through, we need to be a bit more careful in our definition of the matrix $A_{((p_1^{(t_1)},\ldots,p_1^{(t_r)},C_1),(p_2^{(t_1)},\ldots,p_2^{(t_r)},C_2))}$. Namely, as done in Sections 2.2 and 2.3, we want to split the variables $(p_1^{(t_1)}, \ldots, p_1^{(t_r)}, C_1)$ and $(p_2^{(t_1)}, \ldots, p_2^{(t_r)}, C_2)$ of the underlying hyperedges *evenly* across the row and column sets $(S_1, \ldots, S_r, S)$ and $(T_1, \ldots, T_r, T)$. However, each element $p^{(t_z)}$ comes with a "weight" of t_z because $p^{(t_z)}$ corresponds to a set of t_z original variables $[n]$. Because we "canceled out" a variable in P_s, the total weight of the remaining elements in each hyperedge is $q - s$. Ideally, we would like to achieve an even split of $(\frac{q-s}{2}, \frac{q-s}{2})$, but this is not always possible.

Fortunately, for the row pruning step to go through, we do not require an exactly even split: we just need that each side of the split has total weight at most $\frac{q}{2}$. A simple greedy algorithm shows that this is in fact always possible, and so we can make the row pruning step go through. This allows us to prove a lower bound of $n \geq \tilde{\Omega}(k^{\frac{q}{q-2}})$ bound for all odd q, finishing the proof.

2.4 A simpler proof with bipartite Kikuchi graphs

The above proof for odd q is substantially more complicated compared to the fairly simple proof for even q sketched in Section 2.1. Recall that this is the case because the proof in Section 2.2 uses the "Cauchy–Schwarz trick" to derive new constraints of arity $2(q-1)$ that have *correlated* randomness. When we group the derived constraints so that they have independent randomness, i.e., we write $A = \sum_{i \in L} b_i A_i$, where each b_i is an independent bit, the success of the row pruning step is dictated by the structure of the derived constraints that contribute to the matrix A_i. However, the derived constraints look like (C, C') where there exists $u \in [n]$ such that $(u, C) \in H_i$ and $(u, C') \in \cup_{j \in R} H_j$, and so not only are they no longer matchings, they can have highly irregular structure. Because of this, we then decompose the original hypergraph matchings $H_1, \ldots, H_k$ into new hypergraphs where the union, over $i \in [k]$ of the new hypergraphs in each "piece" of the decomposition is regular. In the case of $q = 3, 5$, this can be handled with a simple decomposition step as done (or could have been done for $q = 5$) in [AGKM23], and for $q \geq 7$ a more involved recursive hypergraph decomposition step along with a partite Kikuchi matrix, as sketched in Section 2.3 and Definition 2.3, is needed.

The key point here is that original hypergraphs $H_1, \dots, H_k$ are matchings, i.e., they are already "regular", and so the additional complexity in the proof for odd q comes from the use of the Cauchy–Schwarz trick, which is the step that derives new constraints that are no longer "regular". If we could somehow avoid using the Cauchy–Schwarz trick entirely, we would never lose the "regularity property", and so we could potentially obtain a substantially simpler and more direct proof that avoids any hypergraph decomposition steps.

Unfortunately, we are not quite able to achieve this goal — in the "top level" case, i.e., when the hypergraph matchings $H_1, \dots, H_k$ already satisfy the global property that $\cup_{i\in[k]} H_i$ has no heavy sets of size s for $2 \le s \le \frac{q+1}{2}$, we still do the Cauchy–Schwarz trick to construct the Kikuchi matrix as sketched in Section 2.2. This means that we will do *one* step of hypergraph decomposition to produce, for each $2 \le s \le \frac{q+1}{2}$ a decomposed instance with hypergraph matchings $H_1^{(s)}, \dots, H_k^{(s)}$.

Now, recall that the reason we required the more complicated recursive hypergraph decomposition step in Section 2.3 is because, if we were to apply the Cauchy–Schwarz trick again to each decomposed instance $H_1^{(s)}, \dots, H_k^{(s)}$, the resulting derived constraints may again not be regular. However, the hypergraphs $H_1^{(s)}, \dots, H_k^{(s)}$ are themselves still matchings, i.e., they are regular, as this property is inherited from the original hypergraphs $H_1, \dots, H_k$. Our key technical contribution, as we now explain, is the introduction of a *bipartite* Kikuchi graph that allows us to refute each decomposed instance $H_1^{(s)}, \dots, H_k^{(s)}$ *without* using the Cauchy–Schwarz trick at all. Because we do not apply the Cauchy–Schwarz trick, our hypergraphs remain matchings, and so we do not need to do a recursive hypergraph decomposition as done in our other proof (Section 2.3). The thresholds that we use in the decomposition step are the original thresholds $d_s = n^{-\frac{2s}{q}+\frac{3}{q}}k$, which we note are a factor of $n^{1/q}$ larger than the thresholds that are needed for our other proof in Section 2.3.

Refuting the decomposed instances with bipartite Kikuchi graphs. Instead of doing the Cauchy–Schwarz trick, we introduce a bipartite Kikuchi graph that is imbalanced. This is perhaps a counterintuitive approach to try, as typically imbalanced matrices do not give good spectral refutations. For example, as explained at the beginning of Section 2.2, one can define an imbalanced matrix (Definition 2.2) that cleanly handles the case of odd q with no hypergraph decomposition steps at all, but the imbalance of the matrix produces the weaker bound of $n \ge \tilde{\Omega}(k^{\frac{q+1}{q-1}})$.

Recall that for each $i \in [k]$, a hyperedge in $H_i^{(s)}$ has the form (C, p) where $p \in P_s$ and $|C| = q - s$. We now define our bipartite Kikuchi graph $A_{C,p}$.

Definition 2.4 (Our imbalanced bipartite Kikuchi graph)**.** For a set $C \in \binom{[n]}{q-s}$ and $p \in P_s$, let $A_{C,p}$ be the adjacency matrix of the following graph. The left vertices are pairs of sets (S_1, S_2) where $S_1 \subseteq [n]$ has size ℓ and $S_2 \subseteq P_s$ has size ℓ as well. The right vertices are pairs of sets $T_1 \subseteq [n]$ and $T_2 \subseteq P_s$, where $|T_1| = \ell + 1 - s$ and $|T_2| = \ell + 1$. We put an edge $((S_1, S_2), (T_1, T_2))$ if $S_1 \oplus T_1 = C$, which implies that $|S_1 \cap C| = \frac{q-1}{2}$ and $|T_1 \cap C| = \frac{q+1}{2} - s$, and also $S_2 \oplus T_2 = \{p\}$, which implies that $p \notin S_2$ and $p \in T_2$.

Definition 2.4 is inspired by our partite matrix (Definition 2.3), as for each row/column, we have a subset for each "variable set", i.e., $[n]$ and P_s. However, unlike Definition 2.3, we now have $|S_1| \neq |T_1|$ and $|S_2| \neq |T_2|$, and this makes the graph $A_{C,p}$ quite imbalanced. The size of the left vertex set is $N_L = \binom{n}{\ell}\binom{|P_s|}{\ell}$, where $|P_s| \approx nk/d_s$. This bound on $|P_s|$ follows because $\cup_{i\in[k]} H_i$ has at

most nk hyperedges in total and each $p \in P_s$ is contained in at least d_s hyperedges. On the other hand, the size of the right vertex set is $N_R = \binom{n}{\ell-1+s}\binom{|P_s|}{\ell+1} \approx \frac{|P_s|}{\ell}\left(\frac{\ell}{n}\right)^{s-1} N_L$. Recall that $\ell = n^{1-\frac{2}{q}}$ and $d_s = n^{-\frac{2s}{q}+\frac{3}{q}}k$, so $N_R \approx n^{1/q} N_L$.

Remark 2.5. We note that $A_{C,p}$ is not the first use of an imbalanced bipartite Kikuchi graph, as imbalanced Kikuchi graphs are used in [Yan24, KM24b]. However, in those works one can easily produce an equivalent Kikuchi graph (i.e., non-bipartite and balanced) with essentially the same properties via one application of the Cauchy–Schwarz derivation trick to the underlying hyperedges. Here, as discussed in detail in Section 2.3, such a strategy fails. We thus view our bipartite Kikuchi graph $A_{C,p}$ as being *inherently* bipartite, as we are unable to construct an equivalent (balanced) Kikuchi graph with analogous properties to it by first deriving constraints on the underlying hypergraph and then forming a balanced Kikuchi graph similar to Definition 2.1 using the derived constraints.

With the graph $A_{C,p}$ defined, we then let $A_i = \sum_{(C,p)\in H_i^{(s)}} A_{C,p}$ and $A = \sum_{i=1}^{k} b_i A_i$. Crucially, because we have not used the Cauchy–Schwarz trick, A_i depends only on $H_i^{(s)}$ and not on any of the other hypergraphs. Because $H_i^{(s)}$ is a matching (over the larger vertex set $[n]$ and P_s), the row pruning calculation is much more straightforward.

Row pruning for the imbalanced bipartite Kikuchi graph. Unlike the case of normal Kikuchi graphs, we now need to argue concentration of both the left and right vertices. The case of the left vertices is rather straightforward: because $H_i^{(s)}$ is a matching and for an edge in $A_{C,p}$, the left vertex contains $\frac{q-1}{2}$ elements of C, which is the subset of "original variables" $[n]$, a similar calculation to the row pruning argument in Section 2.1 shows concentration of the degrees of the left vertices provided that the average left degree d_L is at least $\Omega(1)$. The average left degree is $d_L \approx \left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} n$, which is roughly $n^{1/q} \gg 1$ since $\ell = n^{1-2/q}$.

The calculation for the right vertices is more interesting. We again use that $H_i^{(s)}$ is a matching to argue concentration provided that the average right degree d_R is at least $\Omega(1)$. Now, we have $d_R \approx \left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s}\left(\frac{\ell}{|P_s|}\right)\cdot n$. As shown earlier, $|P_s| \le nk/d_s \approx n^{1+\frac{2s}{q}-\frac{3}{q}}$, using our threshold for d_s. Substituting in $\ell = n^{1-2/q}$ and the above bound on $|P_s|$, we see that $d_R \ge \Omega(1)$ holds. Thus, the row pruning step goes through, and we are able to prove the $n \ge \tilde{\Omega}(k^{\frac{q}{q-2}})$ bound with a substantially simpler proof compared to our proof sketch in Section 2.3.

Our bipartite Kikuchi graph compared to the naive imbalanced matrix. Why does the matrix $A_{C,p}$ succeed in yielding a $k^{\frac{q}{q-2}}$ lower bound, whereas the naive ℓ vs. $\ell+1$ matrix (Definition 2.2) only yields a $k^{\frac{q+1}{q-1}}$ lower bound? Recall that for an edge $((S_1,S_2),(T_1,T_2))$ in the matrix $A_{C,p}$, we have split C across S_1 and T_1 as $|S_1 \cap C| = \frac{q-1}{2}$ and $|T_1 \cap C| = \frac{q+1}{2} - s$. Because $p \notin S_2$ and $p \in T_2$, this means that the row set contains $\frac{q-1}{2}$ variables from (C,p), and the column set "effectively" contains $\frac{q+1}{2}$ variables from (C,p): it has $\frac{q+1}{2}-s$ variables contained in T_1 and then an extra s from T_2 because p_s as a variable "represents" a set of size s. Notice that the $\frac{q-1}{2}$ vs. $\frac{q+1}{2}$ split is precisely the split used by the ℓ vs. $\ell+1$ matrix. So, it is reasonable to ask: why is the matrix $A_{C,p}$ performing better?

The reason lies in the fact that the chosen threshold is $d_s = \left(\frac{\ell}{n}\right)^{s-\frac{1}{2}} k = n^{-\frac{2s}{q}+\frac{3}{q}}k$ has an "extra

factor" of $n^{1/q}$ when viewed from the following perspective. Recall that the thresholds d_s are used only for $2 \le s \le \frac{q+1}{2}$. But, if we substitute in $s = 1$, we get $d_1 = n^{1/q}k$, which we can view as giving us a bound on the maximum degree of a singleton v that we are able to tolerate. However, in the hypergraph $H := \cup_{i=1}^{k} H_i$, each singleton v has $\deg_H(v) \le k$, as H is a union of matchings. So, d_s is a $n^{1/q}$ factor "larger" than we might expect. In fact, following intuition from [GKM22], we would like $\deg(Q)$ to "drop" by a factor of ℓ/n per additional vertex included into the set Q, i.e., we intuitively set $d'_1 = k$ and then take $d'_{s+1} = (\ell/n)d'_s$ for larger s. This yields the threshold $d'_s = \left(\frac{\ell}{n}\right)^{s-1} k = n^{-\frac{2s}{q}+\frac{2}{q}} k$, which is the threshold we used in Section 2.3. However, the threshold d'_s is *not* the threshold that arises out of the approach of [AGKM23] (Section 2.2); we can tolerate an extra factor of $n^{1/q}$.

Let us now explain why this $n^{1/q}$ factor is critical. We expect the left/right degrees to be about $\left(\frac{\ell}{n}\right)^{\#\text{ variables in split}} \cdot n$, i.e., $\left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \cdot n$ for the left degree and $\left(\frac{\ell}{n}\right)^{\frac{q+1}{2}} \cdot n$ for the right degree. If this happens, then we must take $\ell = n^{1-2/(q+1)}$ rather than $n^{1-2/q}$. This results in the weaker $k^{\frac{q+1}{q-1}}$ bound, and is precisely what happens with the naive matrix from Definition 2.2. However, for the matrix $A_{C,p}$, the fact that d_s has an extra factor of $n^{1/q}$ boosts the average right degree by a factor of $n^{1/q}$. This means that our right degree is roughly $\left(\frac{\ell}{n}\right)^{\frac{q+1}{2}} \cdot n^{1+1/q}$, and this allows us to still take $\ell = n^{1-2/q}$. Notice that when we computed the sizes of the left and right vertex sets for $A_{C,p}$, we had $N_R \approx n^{1/q} N_L$, whereas in the naive imbalanced matrix of Definition 2.2 one has $N_R \approx n^{2/q} N_L$.

Postmortem: the power of the bipartite Kikuchi matrix. The proof of the $n \ge \tilde{\Omega}(k^{\frac{q}{q-2}})$ lower bound using the bipartite Kikuchi matrices sketched above is substantially simpler than our other proof (sketched in Section 2.3) that follows the more well-trodden path of "hypergraph decomposition + Cauchy–Schwarz" used in prior works ([GKM22, HKM23, AGKM23, KM24a, Yan24, KM24b]). The success of the bipartite matrix in this setting comes as quite a surprise to us, as it is contrary to the conventional wisdom that imbalanced matrices yield poorer spectral certificates compared to balanced matrices. Moreover, the simplicity of the analysis is not just nice for aesthetic reasons: Theorem 1 obtains a better dependence on $\log k$, δ, and ε in the case of $q = 3$ as compared to the lower bound of [AGKM23]. In fact, the $\log k$ dependence in Theorem 1 is exactly the same as the dependence obtained for even q, and the loss in the δ and ε factors comes from the "top level" instance where we still use the Cauchy–Schwarz trick.

Roadmap. The full proof of Theorem 1 is presented in Sections 4 to 6; we give preliminaries and notation in Section 3. In Section 4, we handle the setup and the hypergraph decomposition step (Section 4.1). In Section 5, we refute the "top level" instance using the Cauchy–Schwarz trick, and in Section 6 we use our new bipartite Kikuchi matrices to refute the subinstances $H_1^{(s)}, \dots, H_k^{(s)}$ for all $2 \le s \le \frac{q+1}{2}$.

3 Preliminaries

3.1 Basic notation and hypergraphs

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.

Definition 3.1 (Hypergraphs and hypergraph matchings)**.** A hypergraph H with vertices $[n]$ is a collection of subsets $C \subseteq [n]$ called hyperedges. We say that a hypergraph H is *q-uniform* if $|C| = q$ for all $C \in H$, and we say that H is a *matching* if all the hyperedges in H are disjoint. For a subset $Q \subseteq [n]$, we define the degree of Q in H, denoted $\deg_H(Q)$, to be $|\{C \in H : Q \subseteq C\}|$.

Definition 3.2 (Bipartite hypergraphs)**.** A bipartite hypergraph H has two vertex sets $[n]$ and P and is a collection of pairs (C, p) with $C \subseteq [n]$ and $p \in P$ called hyperedges. We say that a bipartite hypergraph H is *q-uniform* if $|C| = q - 1$ for all $(C, p) \in H$, and we say that H is a *matching* if all the hyperedges in H are disjoint. That is, for (C, p) and (C', p') in H, it holds that $C \cap C' = \emptyset$ and $p \neq p'$.

3.2 Locally decodable codes

We refer the reader to the survey [Yek12] for background.

A code is typically defined as a map $C \colon \{0,1\}^k \to \{0,1\}^n$. However, for our proofs it will be more convenient to view a code as taking values in $\{-1, 1\}$ rather than $\{0, 1\}$; we switch between the two notations via the map $0 \mapsto 1$ and $1 \mapsto -1$. For a code $C \colon \{-1,1\}^k \to \{-1,1\}^n$, we will write $x \in C$ to denote an $x = C(b)$ for some $b \in \{0,1\}^k$.

A locally decodable code is a code where one can recover any bit b_i of the original message b with good confidence while only reading a few bits of the encoded string in the presence of errors.

Definition 3.3 (Locally Decodable Code)**.** A code $C \colon \{-1,1\}^k \to \{-1,1\}^n$ 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 $y \in \{-1,1\}^n$, 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 $b \in \{-1,1\}^k$, $i \in [k]$, and all $y \in \{-1,1\}^n$ such that $\Delta(y, C(b)) \leq \delta n$, $\Pr[\mathrm{Dec}^y(i) = b_i] \geq \frac{1}{2} + \varepsilon$. 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 3.4 (Normal LDC)**.** A code $C \colon \{-1,1\}^k \to \{-1,1\}^n$ is (q, δ, ε)-normally decodable if for each $i \in [k]$, there is a q-uniform hypergraph matching H_i with at least δn hyperedges such that for every $C \in H_i$, it holds that $\Pr_{b \leftarrow \{-1,1\}^k}[b_i = \prod_{v \in C} C(b)_v] \geq \frac{1}{2} + \varepsilon$.

Fact 3.5 (Reduction to LDC Normal Form, Lemma 6.2 in [Yek12])**.** *Let $C\colon \{-1,1\}^k \to \{-1,1\}^n$ be a code that is (q,δ,ε)-locally decodable. Then, there is a code $C'\colon \{-1,1\}^k \to \{-1,1\}^{O(n)}$ that is (q,δ',ε') normally decodable, with $\delta' \geq \varepsilon\delta/3q^2 2^{q-1}$ and $\varepsilon' \geq \varepsilon/2^{2q}$.*

3.3 Matrix concentration inequalities

We will make use of the following non-commutative Khintchine inequality [LP91].

Fact 3.6 (Rectangular Matrix Khintchine Inequality, Theorem 4.1.1 of [Tro15])**.** *Let $X_1,\ldots,X_k$ be fixed $d_1 \times d_2$ matrices and $b_1,\ldots,b_k$ 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*

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

3.4 Binomial coefficient inequalities

In this section, we state and prove the following fact about binomial coefficients that we will use.

Fact 3.7. *Let n,ℓ,q be positive integers with $\ell \leq n$. Let q be constant and ℓ, n be asymptotically large with $\ell = o(n)$. Then,*

$$\frac{\binom{n}{\ell-q}}{\binom{n}{\ell}} = \Theta\left(\left(\frac{\ell}{n}\right)^q\right),$$
$$\frac{\binom{n-q}{\ell}}{\binom{n}{\ell}} = \Theta(1)\,.$$

Proof. We have that

$$\frac{\binom{n}{\ell-q}}{\binom{n}{\ell}} = \frac{\binom{\ell}{q}}{\binom{n-\ell+q}{q}}\,.$$

Using that $\left(\frac{a}{b}\right)^b \leq \binom{a}{b} \leq \left(\frac{ea}{b}\right)^b$ finishes the proof of the first equation.

We also have that

$$\frac{\binom{n-q}{\ell}}{\binom{n}{\ell}} = \frac{(n-q)!(n-\ell)!}{n!(n-\ell-q)!} = \prod_{i=0}^{q-1} \frac{n-\ell-i}{n-i} = \prod_{i=0}^{q-1}\left(1-\frac{\ell}{n-i}\right),$$

and this is $\Theta(1)$ since $\ell = o(n)$ and q is constant. □

4 Proof of Theorem 1

In this section, we begin the proof of Theorem 1. By Fact 3.5, we may assume that we start with a code C in normal form. Namely, C is map $C\colon \{-1,1\}^k \to \{-1,1\}^n$, and there exist q-uniform hypergraphs

$H_1, \dots, H_k$ of size exactly δn such that for every $C \in H_i$, it holds that $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[b_i \prod_{v \in C} C(b)_v] \geq \varepsilon$. We will show that $k \leq O(n^{1-\frac{2}{q}} \delta^{-2-\frac{2}{q}} \varepsilon^{-4})$, which implies Theorem 1.

To begin, we let $\Phi_b(x)$ denote the following polynomial:

$$\Phi_b(x) := \sum_{i=1}^{k} \sum_{C \in H_i} b_i \prod_{v \in C} x_v \,.$$

Because $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[b_i \prod_{v \in C} C(b)_v] \geq \varepsilon$, it follows that $\mathbb{E}_b[\Phi_b(C_b(x))] \geq \varepsilon \sum_{i=1}^{k} |H_i| \geq \varepsilon \delta n k$. Hence, we have that $\mathbb{E}_b[\mathrm{val}(\Phi_b)] \geq \varepsilon \delta n k$, where $\mathrm{val}(\Phi_b) := \max_{x \in \{-1,1\}^n} \Phi_b(x)$.

Overview: refuting the q-XOR instance Φ_b. It thus remains to bound $\mathbb{E}_b[\mathrm{val}(\Phi_b)]$. We will do this by building on the spectral methods of [GKM22, AGKM23]. As discussed in Sections 2.3 and 2.4, the argument proceeds in two steps.

(1) **Hypergraph decomposition:** First, we decompose the hypergraphs $H_1, \dots, H_k$ informally as follows. For $2 \leq t \leq \frac{q+1}{2}$, we define "degree thresholds" d_t where $d_t := \left(\frac{\ell}{n}\right)^{t-\frac{3}{2}} k$. Then, for every $Q \subseteq [n]$ of size s with $2 \leq s \leq \frac{q+1}{2}$, we call Q "heavy" if Q is contained in more than d_s hyperedges in the multiset $\cup_{i=1}^{k} H_i$. For each heavy Q, we introduce a new variable y_{p_Q} and let P_s be the set of the labels p_Q corresponding to $|Q| = s$. Then, for each hyperedge C, if $Q \subseteq C$ is the largest heavy Q contained in C, we replace the hyperedge C with $(C \setminus Q, p_Q)$, where $p_Q \in P_s$. We thus produce, for each $i \in [k]$, hypergraphs $H_i^{(s)}$ for each $2 \leq s \leq \frac{q+1}{2}$ where a hyperedge in $H_i^{(s)}$ has the form (C', p) for some $p \in P_s$ and $C' \subseteq [n]$ with $|C'| = q - s$, along with the hypergraph H_i' of "leftover edges".

(2) **Refutation:** With the decomposition in hand, we then produce polynomials $\Psi_b^{(s)}$ for each $2 \leq s \leq \frac{q+1}{2}$, along with a polynomial Ψ_b, such that $\Psi_b + \sum_{s=2}^{\frac{q+1}{2}} \Psi_b^{(s)} = \Phi_b$. We then produce upper bound $\mathbb{E}_b[\mathrm{val}(\Psi_b)]$ as well as $\mathbb{E}_b[\mathrm{val}(\Psi_b^{(s)})]$ for each polynomial $\Psi_b^{(s)}$ in the decomposition. Combining these bounds allows us to upper bound $\mathbb{E}_b[\mathrm{val}(\Phi_b)]$ and finishes the proof. As discussed in Section 2.4, our key technical contribution is designing the matrix whose spectral norm upper bounds $\mathbb{E}_b[\mathrm{val}(\Psi_b^{(s)})]$ for each of the "decomposed" instances $\Psi^{(s)}$ for $2 \leq s \leq \frac{q+1}{2}$.

We now formally describe the decomposition process.

Lemma 4.1 (Hypergraph Decomposition). *Let $H_1, \dots, H_k$ be q-uniform hypergraphs on n vertices, and let H be the multiset $H := \cup_{i=1}^{k} H_i$.*

For each $2 \leq s \leq \frac{q+1}{2}$, let d_s be a positive integer such that $d_2 \geq d_s \geq \cdots \geq d_{\frac{q+1}{2}} \geq 1$, and let $P_s := \{Q \in \binom{[n]}{s} : \deg_H(Q) > d_s\}$. Then, there are q-uniform hypergraphs $H_1', \dots, H_k'$ and, for each $2 \leq s \leq \frac{q+1}{2}$, bipartite hypergraphs $H_1^{(s)}, \dots, H_k^{(s)}$, with the following properties.

(1) Each $H_i^{(s)}$ is a bipartite hypergraph where each hyperedge contains $q - s$ left vertices in $[n]$ and one right vertex $p \in P_s$. Furthermore, $|P_s| \leq O(|H|/d_s)$.

(2) Each H_i' is a subset of H_i.

(3) *For each $i \in [k]$, there is a one-to-one correspondence between hyperedges $C \in H_i$ and the hyperedges in $H'_i, H_i^{(2)}, \ldots, H_i^{(\frac{q+1}{2})}$ given by $(C, p) \in H_i^{(s)} \mapsto C \cup p \in H_i$ and $C \in H'_i \mapsto C \in H_i$.*

(4) *Let $H' := \cup_{i=1}^k H'_i$. Then, for any $Q \in \binom{[n]}{s}$ with $2 \le s \le \frac{q+1}{2}$, it holds that $\deg_{H'}(Q) \le d_s$.*

(5) *If H_i is a matching, then H'_i and $H_i^{(s)}$ for $2 \le s \le \frac{q+1}{2}$ are also matchings.*

The proof of Lemma 4.1 follows by using a simple greedy algorithm, and is given in Section 4.1.

Given the decomposition, the two main technical parts of the proof are given by the following two theorems. In the first theorem, we refute the q-XOR instance resulting from the hypergraph H', and in the second theorem we refute the bipartite $(q-s+1)$-XOR instances from the hypergraphs $H^{(s)}$ for each $2 \le s \le \frac{q+1}{2}$.

Theorem 4.2 (Refuting the regular q-XOR instance)**.** *Let $q \ge 3$ be an odd integer. Let k, n be positive integers and $\delta \in (0,1)$. Let $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$, and suppose that $k \ge 4\ell$. For $2 \le t \le \frac{q+1}{2}$, let $d_t := \left(\frac{\ell}{n}\right)^{t-\frac{3}{2}} k$.*

Let $H_1, \ldots, H_k$ be q-uniform hypergraph matchings on $[n]$ of size $\le \delta n$, and suppose that for every $Q \subseteq [n]$ with $2 \le |Q| \le \frac{q+1}{2}$, it holds that $\deg_H(Q) \le d_{|Q|}$, where $H := \cup_{i=1}^k H_i$. Let $\Psi_b(x)$ be the polynomial in the variable $x_1, \ldots, x_n$ defined as

$$\Psi_b(x) = \sum_{i=1}^{k} \sum_{C \in H_i} b_i \prod_{v \in C} x_v \,.$$

Then, $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\Psi_b)] \le O(n\sqrt{\delta k}) \cdot (k\ell \log n)^{1/4}$.

Theorem 4.3 (Refuting the bipartite instances)**.** *Let $q \ge 3$ be an odd integer, and let $2 \le s \le \frac{q+1}{2}$. Let k, n be positive integers and $\delta \in (0,1)$. Let $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$, and suppose that $k \ge 4\ell$. For $2 \le t \le \frac{q+1}{2}$, let $d_t := \left(\frac{\ell}{n}\right)^{t-\frac{3}{2}} k$. Let $P_s \subseteq \binom{[n]}{s}$ be a set with $4\ell \le |P_s| \le O\left(\frac{nk}{d_s}\right)$.*

Let $H_1^{(s)}, \ldots, H_k^{(s)}$ be bipartite $(q-s+1)$-uniform hypergraph matchings on $\binom{[n]}{q-s} \times P_s$ of size at most δn. Let $\Psi_b^{(s)}(x, y)$ be the polynomial in the variable $x_1, \ldots, x_n$ and $\{y_p\}_{p \in P_s}$ defined as

$$\Psi_b^{(s)}(x, y) = \sum_{i=1}^{k} \sum_{(C,p) \in H_i} b_i y_p \prod_{v \in C} x_v \,.$$

Then, $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\Psi_b^{(s)})] \le \delta n O(\sqrt{k\ell \log n})$.

We prove Theorem 4.2 in Section 5, and we prove Theorem 4.3 in Section 6.

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

Proof of Theorem 1. By Fact 3.5, we may assume that our code C is in LDC normal form, and our goal is to show that $k \le O(\varepsilon^{-4}\delta^{-2}\ell \log n)$ holds, where $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$, which implies Theorem 1. We

will assume that δ satisfies $\delta \geq n^{-\frac{2}{q+2}}$, as otherwise $\delta^{-2}\ell \geq n$ holds, and so $k \leq n \leq O(\varepsilon^{-4}\delta^{-2}\ell \log n)$ trivially holds.

For each $2 \leq s \leq \frac{q+1}{2}$, define $d_s := \left(\frac{\ell}{n}\right)^{s-\frac{3}{2}} k$. We will assume that $k \geq 4\ell$, as otherwise we are already done. Because of this assumption, we have $\left(\frac{\ell}{n}\right)^{\frac{q}{2}-1} k \geq 1$, which implies that $d_s \geq 1$ for all $2 \leq s \leq \frac{q+1}{2}$. We can thus apply Lemma 4.1 with these thresholds, which decomposes each H_i into H'_i and $H_i^{(s)}$ for $2 \leq s \leq \frac{q+1}{2}$. Note that $|P_s| \leq O(\delta n k / d_s) = O(1) \cdot \left(\frac{n}{\ell}\right)^{s-\frac{3}{2}} \cdot \delta n$.

The one-to-one correspondence property in Lemma 4.1 implies that for each $b \in \{-1,1\}^k$ and every $x \in \{-1,1\}^n$, if we set $y_p = \prod_{v \in p} x_v$ for each $p \in P_s$ and $2 \leq s \leq \frac{q+1}{2}$, then it holds that $\Phi_b(x) = \Psi_b(x) + \sum_{s=2}^{\frac{q+1}{2}} \Psi_b^{(s)}(x,y)$.

We can now apply Theorems 4.2 and 4.3 to bound $\mathbb{E}[\mathrm{val}(\Psi_b)]$ and $\mathbb{E}[\mathrm{val}(\Psi_b^{(s)})]$. However, it is possible that the condition that $|P_s| \geq 4\ell$ does not hold. But, if $|P_s| \leq 4\ell$, then the conclusion of Theorem 4.3 still holds. This is because for any b, $\mathrm{val}(\Psi_b^{(s)}) \leq \sum_{i=1}^k |H_i^{(s)}|$ trivially holds, and we also have $\sum_{i=1}^k |H_i^{(s)}| \leq |P_s| d_s$, as each $p \in P_s$ contributes at most d_s hyperedges to $\cup_{i=1}^k H_i^{(s)}$. Hence, $\mathrm{val}(\Psi_b^{(s)}) \leq \ell d_s$ in this case, which is at most $\delta n O(\sqrt{k\ell \log n})$ when $\delta = \Omega(n^{-\frac{2}{q+2}})$.

We thus have that

$$\varepsilon\delta n k \leq \mathbb{E}[\mathrm{val}(\Phi_b)] \leq \mathbb{E}[\mathrm{val}(\Psi_b)] + \sum_{s=2}^{\frac{q+1}{2}} \mathbb{E}[\mathrm{val}(\Psi_b^{(s)})] \leq O(n\sqrt{\delta k}) \cdot (k\ell \log n)^{1/4} + \delta n O(\sqrt{k\ell \log n}).$$

We have two cases. If $O(n\sqrt{\delta k}) \cdot (k\ell \log n)^{1/4}$ is larger than $\delta n O(\sqrt{k\ell \log n})$, then we conclude that

$$\varepsilon\delta n k \leq O(n\sqrt{\delta k}) \cdot (k\ell \log n)^{1/4} \implies k \leq O(\varepsilon^{-4}\delta^{-2}\ell \log n),$$

and if $O(n\sqrt{\delta k}) \cdot (k\ell \log n)^{1/4}$ is smaller than $\delta n O(\sqrt{k\ell \log n})$, we conclude that

$$\varepsilon\delta n k \leq \delta n O(\sqrt{k\ell \log n}) \implies k \leq O(\varepsilon^{-1}\ell \log n).$$

Thus, we have $k \leq O(\varepsilon^{-4}\delta^{-2}\ell \log n) = O(n^{1-\frac{2}{q}}\delta^{-2-\frac{2}{q}}\varepsilon^{-4} \log n)$, which finishes the proof. □

4.1 Hypergraph decomposition: proof of Lemma 4.1

We prove Lemma 4.1 by analyzing the following greedy algorithm.

Algorithm 4.4.

Given: q-uniform hypergraphs $H_1, \ldots, H_k$ and parameters $d_2 \geq d_3 \geq \cdots \geq d_{\frac{q+1}{2}} \geq 1$.

Output: q-uniform hypergraphs $H'_1, \ldots, H'_k$ and for $2 \leq s \leq \frac{q+1}{2}$, bipartite $(q-s+1)$-uniform hypergraphs $H_1^{(s)}, \ldots, H_k^{(s)}$ over the left vertex set $[n]$ and right vertex set $P_s \subseteq \binom{[n]}{s}$.

Operation:

1. **Initialize:** $H'_i = H_i$ for all $i \in [k]$, $P_s = \emptyset$, and $P'_s = \{Q \in \binom{[n]}{s} : \deg_{H'}(Q) > d_s\}$, where $H' = \cup_{i\in[k]} H'_i$.
2. **For** $t = \frac{q+1}{2}, \ldots, 1$**:**
   (1) **While** P'_t **is nonempty:**
      (a) Choose $p \in P'_t$ arbitrarily.
      (b) Choose an arbitrary set of d_t+1 hyperedges in H' containing the set p. Namely, let $C_{i_1}, \ldots, C_{i_{d_t+1}}$ be hyperedges in H' where $C_{i_1} \in H'_{i_1}, \ldots, C_{i_{d_t+1}} \in H'_{i_{d_t+1}}$.
      (c) Add p to P_t and for each $r \in [d_t + 1]$, remove C_{i_r} from H'_{i_r} and add the hyperedge $(C_{i_r} \setminus p, p)$ to $H^{(t)}_{i_r}$.
      (d) Recompute $P'_t = \{Q \in \binom{[n]}{t} : \deg_{H'}(Q) > d_t\}$.
3. Output $H'_1, \ldots, H'_k$ and $H^{(s)}_1, \ldots, H^{(s)}_k$ for all $2 \le s \le \frac{q+1}{2}$.

We now need to show that the output of Algorithm 4.4 has the desired properties.

Item (1) holds by construction, as each $p \in P_s$ has size s so when the hyperedge C is split into $(C \setminus p, p)$, $|C \setminus p| = q - s$. We have that $|P_s| \le O(|H|/d_s)$, as each hyperedge $C \in H$ has (crudely) at most $2^q = O(1)$ subsets of size exactly s, and each $p \in P_s$ must appear at least $d_s + 1$ times across hyperedges in H.

Item (2) holds by construction, as we start with $H'_i = H_i$ and only remove edges from H'_i.

Item (3) holds because each hyperedge $C \in H_i$ is either never removed (in which case it appears in H'_i), or it is removed exactly once. If it is removed by choosing some $p \in P_s$, then it appears in $H^{(s)}_i$ as the hyperedge $(C \setminus p, p)$.

Item (4) holds because otherwise the algorithm would not have terminated.

Item (5) holds because the operations done by Algorithm 4.4 do not affect the matching property. This finishes the proof.

5 Refuting the Regular q-XOR Instance

In this section, we prove Theorem 4.2, which we recall below.

Theorem 4.2 (Refuting the regular q-XOR instance)**.** *Let $q \ge 3$ be an odd integer. Let k, n be positive integers and $\delta \in (0,1)$. Let $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$, and suppose that $k \ge 4\ell$. For $2 \le t \le \frac{q+1}{2}$, let $d_t := \left(\frac{\ell}{n}\right)^{t-\frac{3}{2}} k$.*

Let $H_1, \ldots, H_k$ be q-uniform hypergraph matchings on $[n]$ of size $\le \delta n$, and suppose that for every $Q \subseteq [n]$ with $2 \le |Q| \le \frac{q+1}{2}$, it holds that $\deg_H(Q) \le d_{|Q|}$, where $H := \cup_{i=1}^k H_i$. Let $\Psi_b(x)$ be the polynomial in the variable $x_1, \ldots, x_n$ defined as

$$\Psi_b(x) = \sum_{i=1}^{k} \sum_{C \in H_i} b_i \prod_{v \in C} x_v \,.$$

Then, $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\Psi_b)] \le O(n\sqrt{\delta k}) \cdot (k\ell \log n)^{1/4}$.

The proof of Theorem 4.2 follows the overall blueprint outlined in the work of [AGKM23], as explained in Section 2.2.

Step 1: the Cauchy-Schwarz trick. First, we show that we can relate $\Psi(x)$ to a certain "Cauchy-Schwarzed" polynomial $f_{L,R}(x)$.

Lemma 5.1 (Cauchy-Schwarz Trick). *Let Ψ be as in Theorem 4.2 and let $L, R \subseteq [k]$ be a random partition of $[k]$, i.e., each $i \in [k]$ appears in L with probability $1/2$, independently, and $R = [k] \setminus L$. Let $f_{L,R}(x)$ be the polynomial defined as*

$$f_{L,R}(x) := \sum_{i\in L, j\in R} \sum_{u\in[n]} \sum_{(u,C_1)\in H_i,(u,C_2)\in H_j} b_i b_j \prod_{v\in C_1} x_v \prod_{v\in C_2} x_v \,.$$

Then, it holds that $(q \operatorname{val}(\Psi))^2 \le q\delta n^2 + 4n\mathbb{E}_{(L,R)} \operatorname{val}(f_{L,R})$. In particular, $\mathbb{E}_{b\in\{-1,1\}^k}[q^2 \cdot \operatorname{val}(\Psi)^2] \le q\delta n^2 + 4n\mathbb{E}_{(L,R)}\mathbb{E}_{b\in\{-1,1\}^k}[\operatorname{val}(f_{L,R})]$.

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

$$\begin{aligned}
(q\Psi(x))^2 &= \left(\sum_{u\in[n]} x_u \sum_{i\in[k]} \sum_{(u,C)\in H_i} b_i x_C\right)^2 \le \left(\sum_{u\in[n]} x_u^2\right)\left(\sum_{u\in[n]}\left(\sum_{i\in[k]}\sum_{(u,C)\in H_i} b_i x_C\right)^2\right) \\
&= n \sum_{u\in[n]} \sum_{i,j\in[k]} \sum_{\substack{(u,C_1)\in H_i \\ (u,C_2)\in H_j}} b_i b_j x_{C_1} x_{C_2} = n\left(q\sum_{i\in[k]}|H_i| + \sum_{u\in[n]}\sum_{i,j\in[k],i\neq j}\sum_{\substack{(u,C_1)\in H_i \\ (u,C_2)\in H_j}} b_i b_j x_{C_1} x_{C_2}\right) \\
&= qn\cdot\delta n + 4n\cdot \mathbb{E}_{(L,R)} f_{L,R}(x) \ ,
\end{aligned}$$

where the first equality is because there are q ways to decompose a set $C_i \in H_i$ with $|C_i| = q$ into a pair (u, C) with $|C| = q-1$, the inequality follows by the Cauchy-Schwarz inequality, and the last equality follows because for a pair of hypergraphs H_i and H_j, we have $i \in L$ and $j \in R$ with probability $1/4$. Finally, $\max_{x\in\{-1,1\}^n} \mathbb{E}_{(L,R)} f_{L,R}(x) \le \mathbb{E}_{(L,R)}[\max_{x\in\{-1,1\}^n} f_{L,R}(x)] = \mathbb{E}_{(L,R)} \operatorname{val}(f_{L,R})$. Thus, we have that $q^2 \cdot \operatorname{val}(\Psi)^2 \le q\delta n^2 + 4n \cdot \mathbb{E}_{(L,R)} \operatorname{val}(f_{L,R})$. □

Step 2: defining the Kikuchi matrices. Next, we define the Kikuchi matrices that we will use and relate them to the polynomial $f_{L,R}$.

Definition 5.2. Let $q \ge 3$ be an odd integer and let $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$. Let (u, C_1) be a hyperedge with $|C_1| = q-1$ and let (u, C_2) be a hyperedge with $|C_2| = q-1$. We define the matrix A_{u,C_1,C_2} to be the matrix indexed by pairs of sets (S_1, S_2) where $S_1, S_2 \subseteq [n]$ and $|S_1| = |S_2| = \ell$, where $A_{u,C_1,C_2}((S_1,S_2),(T_1,T_2)) = 1$ if $S_1 \oplus T_1 = C_1$ and $S_2 \oplus T_2 = C_2$, and 0 otherwise. We note that this is equivalent to $|S_1 \cap C_1| = |T_1 \cap C_1| = \frac{q-1}{2}$ and $|S_2 \cap C_2| = |T_2 \cap C_2| = \frac{q-1}{2}$.

We will also view the matrix A_{u,C_1,C_2} as the adjacency matrix of a graph G_{u,C_1,C_2}.

For $i \neq j \in [k]$, we define $A_{i,j} := \sum_{u\in[n]} \sum_{(u,C_1)\in H_i,(u,C_2)\in H_j} A_{u,C_1,C_2}$. We also define $A_i := \sum_{j\in R} b_j A_{i,j}$ and $A := \sum_{i\in L} b_i A_i$.

Claim 5.3. For each (u, C_1, C_2), the matrix A_{u,C_1,C_2} defined in Definition 5.2 satisfies the following properties.

(1) The matrix A_{u,C_1,C_2} has exactly $D = \binom{q-1}{\frac{q-1}{2}}^2 \binom{n-(q-1)}{\ell-\frac{q-1}{2}}^2$ nonzero entries.

(2) For each $x \in \{-1,1\}^n$, let $z \in \{-1,1\}^{\binom{n}{\ell}}$ be defined as: $z_{S_1,S_2} := \prod_{v\in S_1} x_v \prod_{v\in S_2} x_v$. Then, $z^\top A_{u,C_1,C_2} z = D \prod_{v\in C_1} x_v \prod_{v\in C_2} x_v$.

In particular, $z^\top A z = D f_{L,R}(x)$.

As additional notation, we let $N := \binom{n}{\ell}$ and $d = \frac{\delta n k D}{N}$.

Proof. To prove Item (1), we will count the number of edges. By definition, we have an edge $((S_1,S_2),(T_1,T_2))$ in the graph with adjacency matrix A_{u,C_1,C_2} iff $|S_1 \cap C_1| = \frac{q-1}{2}$ and $|S_2 \cap C_2'| = \frac{q-1}{2}$. The number of such S_1 is $\binom{q-1}{\frac{q-1}{2}}\binom{n-(q-1)}{\ell-\frac{q-1}{2}}$, and the number of such S_2 is the same. Hence, Item (1) holds.

To prove Item (2), we observe that

$$z^\top A_{u,C_1,C_2} z = \sum_{((S_1,S_2),(T_1,T_2))\in E(G_{u,C_1,C_2})} z_{S_1,S_2} w_{T_1,T_2} = \sum_{((S_1,S_2),(T_1,T_2))\in E(G_{u,C_1,C_2})} \prod_{v\in S_1} x_v \prod_{v\in S_2} x_v \prod_{v\in T_1} x_v \prod_{v\in T_2} x_v$$

$$= \sum_{((S_1,S_2),(T_1,T_2))\in E(G_{u,C_1,C_2})} \prod_{v\in S_1\oplus T_1} x_v \prod_{v\in S_2\oplus T_2} x_v = \sum_{((S_1,S_2),(T_1,T_2))\in E(G_{u,C_1,C_2})} \prod_{v\in C_1} x_v \prod_{v\in C_2} x_v = D \prod_{v\in C_1} x_v \prod_{v\in C_2} x_v .$$

The "in particular" follows immediately from Item (2) and the definition of A. □

Step 3: finding an approximately regular submatrix. The key technical lemma, which we shall prove in Section 5.1, shows that we can find an approximately biregular subgraph of A_i for each $i \in L$.

Lemma 5.4 (Approximately regular submatrix)**.** *For $i \in L$, let A_i be defined as in Definition 5.2. There exists a positive integer D' with $D \geq D' \geq \frac{D}{2}$ such that the following holds. For each $i \in L$, $(u, C_1) \in H_i$, $j \in R$, and $(u, C_2) \in H_j$, there exists a matrix $B_{i,u,C_1,C_2} \in \{0,1\}^{\binom{n}{\ell}}$ with the following properties:*

(1) B_{i,u,C_1,C_2} is a "subgraph" of A_{u,C_1,C_2}. Namely, $B_{i,u,C_1,C_2} = B^\top_{i,u,C_1,C_2}$ and if $B_{i,u,C_1,C_2}((S_1,S_2),(T_1,T_2)) = 1$, then we also have $A_{u,C_1,C_2}((S_1,S_2),(T_1,T_2)) = 1$.

(2) B_{i,u,C_1,C_2} has exactly D' nonzero entries.

(3) The matrix $B_i := \sum_{j\in R} \sum_{u\in[n]} \sum_{(u,C_1)\in H_i,(u,C_2)\in H_j} B_{i,u,C_1,C_2}$ has at most $O(d)$ nonzero entries per row or column, where $d = \frac{\delta n k D}{N}$.

Step 4: finishing the proof. With Lemma 5.4 in hand, we can now finish the proof. Let B_i be the matrix defined in Lemma 5.4. We observe that Items (1), (2), and (3) in Lemma 5.4, along with Claim 5.3, imply that for each $x \in \{-1,1\}^n$, there exists $z \in \{-1,1\}^N$ such that $z^\top B z = D' f_{L,R}(x)$. Hence, for any $x \in \{-1,1\}^n$, it holds that $\mathrm{val}(f_{L,R}) \leq \|B\|_2 \cdot N$. We thus have that $\mathbb{E}_b[\mathrm{val}(f_{L,R})] \leq \frac{N}{D'} \mathbb{E}_b[\|B\|_2]$.

It remains to bound $\mathbb{E}_b[\|B\|_2]$, which we do using Matrix Khintchine (Fact 3.6) in the following claim.

Claim 5.5. Let B be the matrix defined in Lemma 5.4. Then, $\mathbb{E}_b[\|B\|_2] \le O(d\sqrt{k\ell \log n})$.

We postpone the proof of Claim 5.5 to the end of this section, and finish the proof of Theorem 4.2. We have that

$$\begin{aligned}\mathbb{E}_b[\mathrm{val}(f_{L,R})] &\le \frac{N}{D'}\mathbb{E}_b[\|B\|_2] \le \frac{2N}{D}O(d\sqrt{k\ell \log n}) \\ &= O(1)\cdot \frac{Nd}{D}\sqrt{k\ell \log n} = \delta nk \cdot O(\sqrt{k\ell \log n})\,,\end{aligned}$$

where we recall that $d = \delta nkD/N$. We note that the above holds for *any* choice of the partition $L \cup R = [k]$. Finally, we recall that by Lemma 5.1, we have that

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

We now finish the proof of Claim 5.5.

Proof of Claim 5.5. By Matrix Khintchine (Fact 3.6), we have $\mathbb{E}_b[\|B\|_2] \le O(\sqrt{\sigma^2 \log N})$, where $\sigma^2 = \|\sum_{i=1}^k B_i^2\|$, as B_i is symmetric. Since B_i is symmetric, $\|B_i\|_2$ is bounded by the maximum ℓ_1-norm of a row in this matrix. By construction of B_i, this is $O(d)$. Hence, $\sigma^2 \le k\cdot O(d)^2 = O(kd^2)$.

We can thus set $\sigma^2 = O(kd^2)$ and apply Fact 3.6 to conclude that $\mathbb{E}_b[\|B\|_2] \le O(d\sqrt{k\log N})$. Recall that we have $N = \binom{n}{\ell} \le n^\ell (nk)^\ell \le n^{O(\ell)}$. Hence, $\log N = O(\ell \log n)$, which finishes the proof. □

5.1 Finding an approximately regular subgraph: proof of Lemma 5.4

In this section, we prove Lemma 5.4. We will prove Lemma 5.4 by using the strategy, due to [Yan24], of bounding "conditional first moments". These moment bounds form the main technical component of the argument.

Lemma 5.6 (Conditional first moment bounds)**.** *Fix $i \in L$. For a vertex (S_1, S_2), let $\deg_i(S_1, S_2)$ denote the degree of (S_1, S_2) in A_i.*

Let $(u, C_1) \in H_i$ and $(u, C_2) \in \cup_{j\in R} H_j$. Let μ_{u,C_1,C_2} denote the distribution over vertices that first chooses a uniformly random edge in A_{u,C_1,C_2} and then outputs a random endpoint. Then, it holds that

$$\mathbb{E}_{(S_1,S_2)\sim\mu_{u,C_1,C_2}}[\deg_i(S_1,S_2)] \le 1 + O(1)\left(\frac{\ell}{n}\right)^{q-1}\delta nk\,.$$

Claim 5.7 (Degree bound). Let $d = \frac{\delta nkD}{N}$. Then, we have that $d \ge \Omega(1)\cdot\left(\frac{\ell}{n}\right)^{q-1}\delta nk$ and that $\left(\frac{\ell}{n}\right)^{q-1}\delta nk \ge \Omega(1)$.

Proof of Claim 5.7. Applying Fact 3.7, we have that

$$\frac{\delta n D}{N} = \delta n \cdot \left(\frac{\binom{q-1}{\frac{q-1}{2}}\binom{n-(q-1)}{\ell-\frac{q-1}{2}}}{\binom{n}{\ell}} \right)^2 \geq \Omega(1) \cdot \delta n \cdot \left(\frac{\ell}{n}\right)^{q-1}.$$

Because $\ell = \lfloor n^{1-2/q}\delta^{-2/q} \rfloor$ and $k \geq \ell$, we have $\left(\frac{\ell}{n}\right)^{q-1} \delta n k \geq 1$, which finishes the proof. □

We postpone the proof of Lemma 5.6 to the end of this subsection, and now use it to finish the proof of Lemma 5.4.

Proof of Lemma 5.4 from Lemma 5.6. Fix $i \in [k]$. Let Γ be a constant (to be chosen later), and let $V_i' = \{(S_1, S_2) : \deg_i(S_1, S_2) \leq \Gamma d\}$.

Let $(u, C_1) \in H_i$ and $(u, C_2) \in \cup_{j \in R} H_j$. We let A'_{i,u,C_1,C_2} be the matrix where $A'_{i,u,C_1,C_2}((S_1, S_2), (T_1, T_2)) = A_{u,C_1,C_2}((S_1, S_2), (T_1, T_2))$ if $(S_1, S_2) \in V'$ and $(T_1, T_2) \in V'$, and otherwise $A'_{i,u,C_1,C_2}((S_1, S_2), (T_1, T_2)) = 0$. Namely, we have "zeroed out" all rows and columns of A_{u,C_1,C_2} that are not in V'. Notice that A'_{i,u,C_1,C_2} depends on $i \in L$ because V' does.

The conditional moment bound from Lemma 5.6, combined with the lower bound on d from Claim 5.7 implies that $\mathbb{E}_{(S_1,S_2)\sim\mu_{u,C_1,C_2}}[\deg_i(S_1, S_2)] \leq O(d)$. Hence, applying Markov's inequality, the number of vertices (S_1, S_2) that are adjacent to an edge labeled by (u, C_1, C_2) and have $\deg_i(S_1, S_2) > \Gamma d$ is at most $O(D/\Gamma)$. Hence, there must be at least $D(1 - O(1/\Gamma))$ edges, i.e., nonzero entries, in $A'_{i,u,C,C'}$.

Now, we let $B_{i,u,C,C'}$ be any subgraph of $A'_{i,u,C,C'}$, where $B_{i,u,C,C'}$ has *exactly* $D' = \lfloor D(1 - O(1/\Gamma)) \rfloor$ edges. This can be achieved by simply removing edges if there are too many. By choosing Γ to be a sufficiently large constant, we ensure that $D' \geq D/2$. Note that because $B_{i,u,C,C'}$ is the adjacency matrix of a graph, it is a symmetric matrix.

To prove the third property, we observe that for any vertex (S_1, S_2), the matrix

$$B_i = \sum_{u \in [n]} \sum_{(u,C_1) \in H_i, (u,C_2) \in \cup_{j \in R} H_j} B_{i,u,C_1,C_2}$$

has at most $\Gamma d = O(d)$ nonzero entries in the (S_1, S_2)-th row or column. Indeed, this follows because it is a subgraph of the original graph A_i, and if (S_1, S_2) had degree $> \Gamma d_L$ in A_i then it has degree 0 in B_i. This finishes the proof. □

It remains to prove Lemma 5.6, which we do now.

Proof of Lemma 5.6. Let $(u, C_1) \in H_i$ and $(u, C_2) \in H_j$ for some $j \in R$. We observe that for any $(S_1, S_2) \in \binom{[n]}{\ell} \times \binom{[n]}{\ell}$, the vertex (S_1, S_2) is adjacent to at most one edge labeled by (u, C_1, C_2). Hence, it follows that μ_{u,C_1,C_2} is uniform over pairs of sets (S_1, S_2) such that $|S_1 \cap C_1| = \frac{q-1}{2}$ and $|S_2 \cap C_2| = \frac{q-1}{2}$. Thus,

$$\mathbb{E}_{(S_1,S_2)\sim\mu_{u,C_1,C_2}}[\deg_i(S_1, S_2)] \leq 1 + \frac{1}{D} \sum_{(u',C_1',C_2')} |\{(S_1, S_2) : |S_1 \cap C_1| = |S_1 \cap C_1'| = |S_2 \cap C_2| = |S_2 \cap C_2'| = \frac{q-1}{2}\}|$$

$$\leq 1 + \frac{1}{D} \sum_{\substack{Z_1 \subseteq C_1 \\ Z_2 \subseteq C_2 \\ |Z_1| = |Z_2| = \frac{q-1}{2}}} \sum_{(u', C_1', C_2')} |\{(S_1, S_2) : \substack{Z_1 \subseteq S_1 \\ Z_2 \subseteq S_2}, \substack{|S_1 \cap C_1'| = \frac{q-1}{2} \\ |S_2 \cap C_2'| = \frac{q-1}{2}}\}| \quad \text{(because } Z_1 \subseteq S_1 \text{ implies } |S_1 \cap C_1| \geq \frac{q-1}{2})$$

$$\leq 1 + \frac{1}{D} \sum_{\substack{Z_1 \subseteq C_1 \\ Z_2 \subseteq C_2 \\ |Z_1| = |Z_2| = \frac{q-1}{2}}} \sum_{\substack{Q_1 \subseteq Z_1 \\ Q_2 \subseteq Z_2}} \sum_{\substack{(u', C_1', C_2') \\ Q_1 = C_1' \cap Z_1 \\ Q_2 = C_2' \cap Z_2}} |\{(R_1, R_2) : \substack{|R_1| = \ell - |Z_1| \\ |R_2| = \ell - |Z_2|}, \substack{|R_1 \cap C_1'| = \frac{q-1}{2} - |Q_1| \\ |R_2 \cap C_2'| = \frac{q-1}{2} - |Q_2|}\}| \quad \text{(by taking } R_1 = S_1 \setminus Z_1)$$

$$\leq 1 + \frac{1}{D} \sum_{\substack{Z_1 \subseteq C_1 \\ Z_2 \subseteq C_2 \\ |Z_1| = |Z_2| = \frac{q-1}{2}}} \sum_{\substack{Q_1 \subseteq Z_1 \\ Q_2 \subseteq Z_2}} \sum_{\substack{(u', C_1', C_2') \\ Q_1 = C_1' \cap Z_1 \\ Q_2 = C_2' \cap Z_2}} \binom{q-1-|Q_1|}{\frac{q-1}{2} - |Q_1|} \binom{n}{\ell - |Z_1| - (\frac{q-1}{2} - |Q_1|)} \binom{q-1-|Q_2|}{\frac{q-1}{2} - |Q_2|} \binom{n}{\ell - |Z_2| - (\frac{q-1}{2} - |Q_2|)}$$

$$\leq 1 + \frac{O(1)}{D} \sum_{\substack{Z_1 \subseteq C_1 \\ Z_2 \subseteq C_2 \\ |Z_1| = |Z_2| = \frac{q-1}{2}}} \sum_{\substack{Q_1 \subseteq Z_1 \\ Q_2 \subseteq Z_2}} \sum_{\substack{(u', C_1', C_2') \\ Q_1 = C_1' \cap Z_1 \\ Q_2 = C_2' \cap Z_2}} \binom{n}{\ell - |Z_1| - (\frac{q-1}{2} - |Q_1|)} \binom{n}{\ell - |Z_2| - (\frac{q-1}{2} - |Q_2|)}$$

Recall that $D \coloneqq \binom{q-1}{\frac{q-1}{2}}^2 \binom{n-(q-1)}{\ell - \frac{q-1}{2}}^2$. By Fact 3.7, we have

$$\frac{1}{D} \binom{n}{\ell - |Z_1| - (\frac{q-1}{2} - |Q_1|)} \binom{n}{\ell - |Z_2| - (\frac{q-1}{2} - |Q_2|)} \leq O(1) \cdot \left(\frac{\ell}{n}\right)^{|Z_1| - |Q_1| + |Z_2| - |Q_2|} = O(1) \cdot \left(\frac{\ell}{n}\right)^{(q-1) - |Q_1| - |Q_2|},$$

as $|Z_1| = |Z_2| = \frac{q-1}{2}$.

For sets Q_1, Q_2, we let $\mu(Q_1, Q_2)$ be the number of (u', C_1', C_2') such that $Q_1 \subseteq C_1'$ and $Q_2 \subseteq C_2'$. We then have that

$$\mathbb{E}_{(S_1, S_2) \sim \mu_{u, C_1, C_2}}[\deg_i(S_1, S_2)] \leq 1 + O(1) \sum_{\substack{Z_1 \subseteq C_1 \\ Z_2 \subseteq C_2 \\ |Z_1| = |Z_2| = \frac{q-1}{2}}} \sum_{\substack{Q_1 \subseteq Z_1 \\ Q_2 \subseteq Z_2}} \mu(Q_1, Q_2) \left(\frac{\ell}{n}\right)^{(q-1) - |Q_1| - |Q_2|}.$$

We now show that $\mu(Q_1, Q_2) \leq O(1) \left(\frac{\ell}{n}\right)^{|Q_1| + |Q_2|} \delta n k$ where $0 \leq |Q_1|, |Q_2| \leq \frac{q-1}{2}$. We have several cases.

(1) $|Q_1| = 0$. In this case, we know that (u', C_2') is in H_j for some $j \in R$ with $Q_2 \subseteq C_2'$. We have three subcases.

   (a) $|Q_2| = 0$. Then, there are at most $q \delta n$ choices for $(u', C_1') \in H_i$, as $|H_i| = \delta n$ and we have q choices for the special element u. Furthermore, given u, there are at most k choices for (u', C_2') with $(u', C_2') \in \cup_{j \in R} H_j$, as each H_j is a matching and $|R| \leq k$. Thus, $\mu(Q_1, Q_2) \leq O(\delta n k)$ in this case, which satisfies the desired bound as $|Q_1| + |Q_2| = 0$.

   (b) $|Q_2| = 1$. Then, there are at most qk choices for $(u', C_2') \in \cup_{j \in R} H_j$ with $Q_2 \subseteq C_2'$. Indeed, this is because each H_j is matching, and $|R| \leq k$. As H_i is a matching, there is at most one choice for $(u', C_1') \in H_i$. Hence, $\mu(Q_1, Q_2) \leq O(k)$ in this case, which is $\leq O(1) \left(\frac{\ell}{n}\right) \delta n k$.

(c) $|Q_2| \geq 2$. Then, there are at most $q d_{|Q_2|}$ choices for $(u', C_2') \in \cup_{j \in R} H_j$, by regularity. As before, given (u, C_2), there is at most one choice for $C_1' \in H_i$. Hence, $\mu(Q_1, Q_2) \leq O(d_{|Q_2|})$ in this case. Since $2 \leq |Q_2| \leq \frac{q-1}{2}$, we have that $d_{|Q_2|} = \left(\frac{\ell}{n}\right)^{|Q_2|-\frac{3}{2}} k$, which is at most $\left(\frac{\ell}{n}\right)^{|Q_2|} \delta n k$ since $\left(\frac{\ell}{n}\right)^{\frac{3}{2}} \delta n \geq 1$.

(2) $|Q_1| \geq 1$. Then, there are at most q choices for $(u', C_1') \in H_i$. It then follows that we have determined $|Q_2| + 1$ elements of $(u', C_2') \in \cup_{j \in R} H_j$, namely u' along with Q_2. We have two subcases.

(a) $|Q_2| = 0$. Then, we have determined one element of (u', C_2'), and so we have at most k choices. Thus, $\mu(Q_1, Q_2) \leq O(k)$ in this case, which is at most $O(1) \left(\frac{\ell}{n}\right)^{|Q_1|} \delta n k$, since $|Q_1| \leq \frac{q-1}{2}$.

(b) $|Q_2| \geq 1$. Then, we have determined at least two elements of (u', C_2'). As $|Q_2| \leq \frac{q-1}{2}$, we have that $|Q_2| + 1 \leq \frac{q+1}{2}$, and so we have at most $d_{|Q_2|+1}$ choices in this case. Thus, $\mu(Q_1, Q_2) \leq O(d_{|Q_2|+1})$. As $d_{|Q_2|+1} = \left(\frac{\ell}{n}\right)^{|Q_2|-\frac{1}{2}} k \leq \left(\frac{\ell}{n}\right)^{\frac{q-1}{2}+|Q_2|} \delta n k \leq \left(\frac{\ell}{n}\right)^{|Q_1|+|Q_2|} \delta n k$, where we use that $|Q_1| \leq \frac{q-1}{2}$, we again have the desired bound on $\mu(Q_1, Q_2)$.

We have thus shown that $\mu(Q_1, Q_2) \leq O(1) \left(\frac{\ell}{n}\right)^{|Q_1|+|Q_2|} \delta n k$. Hence,

$$\mathbb{E}_{(S_1,S_2)\sim\mu_{u,C_1,C_2}}[\deg_i(S_1, S_2)] \leq 1 + O(1) \sum_{\substack{Z_1 \subseteq C_1 \\ Z_2 \subseteq C_2 \\ |Z_1|=|Z_2|=\frac{q-1}{2}}} \sum_{\substack{Q_1 \subseteq Z_1 \\ Q_2 \subseteq Z_2}} \left(\frac{\ell}{n}\right)^{|Q_1|+|Q_2|} \delta n k \cdot \left(\frac{\ell}{n}\right)^{(q-1)-|Q_1|-|Q_2|}$$

$$\leq 1 + O(1) \left(\frac{\ell}{n}\right)^{q-1} \delta n k \,,$$

which finishes the proof. □

6 Refuting the Bipartite Instances

In this section, we prove Theorem 4.3, which we recall below.

Theorem 4.3 (Refuting the bipartite instances)**.** *Let $q \geq 3$ be an odd integer, and let $2 \leq s \leq \frac{q+1}{2}$. Let k, n be positive integers and $\delta \in (0, 1)$. Let $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$, and suppose that $k \geq 4\ell$. For $2 \leq t \leq \frac{q+1}{2}$, let $d_t := \left(\frac{\ell}{n}\right)^{t-\frac{3}{2}} k$. Let $P_s \subseteq \binom{[n]}{s}$ be a set with $4\ell \leq |P_s| \leq O\left(\frac{nk}{d_s}\right)$.*

Let $H_1^{(s)}, \ldots, H_k^{(s)}$ be bipartite $(q - s + 1)$-uniform hypergraph matchings on $\binom{[n]}{q-s} \times P_s$ of size at most δn. Let $\Psi_b^{(s)}(x, y)$ be the polynomial in the variable $x_1, \ldots, x_n$ and $\{y_p\}_{p \in P_s}$ defined as

$$\Psi_b^{(s)}(x, y) = \sum_{i=1}^{k} \sum_{(C,p) \in H_i} b_i y_p \prod_{v \in C} x_v \,.$$

Then, $\mathbb{E}_{b \leftarrow \{-1,1\}^k}[\mathrm{val}(\Psi_b^{(s)})] \leq \delta n O(\sqrt{k\ell \log n})$.

For notational simplicity, we will assume that $n^{1-2/q} \cdot \delta^{-2/q}$ is an integer, so that $\ell = n^{1-2/q} \cdot \delta^{-2/q}$. We note that if $n^{1-2/q} \cdot \delta^{-2/q}$ is not an integer, then we can set $\ell = \lfloor n^{1-2/q} \cdot \delta^{-2/q} \rfloor$, and this only changes the bounds of the following proof by an $O(1)$-factor. We will also write H_i instead of $H_i^{(s)}$ to simplify notation.

We begin by defining the following Kikuchi matrix.

Definition 6.1 (Kikuchi matrix for bipartite hypergraphs)**.** For each $C \in \binom{[n]}{q-s}$ and $p \in P_s$, we define the bipartite graph $G_{C,p}$, parametrized by ℓ, as follows. The left vertex set V_L is the set of pairs of sets (S_1, S_2) where $S_1 \in \binom{[n]}{\ell}$ and $S_2 \in \binom{P_s}{\ell}$. The right vertex set V_R is the set of pairs of sets (T_1, T_2) where $T_1 \in \binom{[n]}{\ell+1-s}$ and $T_2 \in \binom{P_s}{\ell+1}$. We add an edge $((S_1, S_2), (T_1, T_2))$, which we view as "labeled" by C, if the following conditions hold:

(1) $S_1 \oplus T_1 = C$ and $S_2 \oplus T_2 = \{p\}$;

(2) $|S_1 \cap C| = \frac{q-1}{2}$ (and so $|T_1 \cap C| = \frac{q+1}{2} - s$).

We can naturally view the graph $G_{C,p}$ as corresponding to its bipartite adjacency matrix $A_{C,p} \in \{0,1\}^{V_L \times V_R}$. For each $i \in [k]$, we define the matrix $A_i = \sum_{(C,p) \in H_i} A_{C,p}$. We let $A = \sum_{i=1}^{k} b_i A_i$.

Claim 6.2. For each $C \in \binom{[n]}{q-s}$ and $p \in P_s$, the matrix $A_{C,p}$ defined in Definition 6.1 satisfies the following properties.

(1) The matrix $A_{C,p}$ has exactly $D = \binom{q-s}{\frac{q-1}{2}}\binom{n-(q-s)}{\ell-\frac{q-1}{2}}\binom{|P_s|-1}{\ell}$ nonzero entries.

(2) For each $x \in \{-1,1\}^n$ and $y \in \{-1,1\}^{P_s}$, let $z \in \{-1,1\}^{V_L}$ and $w \in \{-1,1\}^{V_R}$ be defined as: $z_{S_1,S_2} := \prod_{v \in S_1} x_v \prod_{p \in S_2} y_p$ and $w_{T_1,T_2} := \prod_{v \in T_1} x_v \prod_{p \in T_2} y_p$. Then, $z^\top A_{C,p} w = D y_p \prod_{v \in C} x_v$.

In particular, $z^\top A w = D\Psi(x, y)$.

As additional notation, we let $N_L := |V_L|$, $N_R := |V_R|$. Finally, we let d_L and d_R denote (upper bounds on) the average left and right degrees of each A_i, i.e., $d_L = \frac{\delta n D}{N_L}$ and $d_R = \frac{\delta n D}{N_R}$.

Proof. To prove Item (1), we will count the number of edges. Let $C \in \binom{[n]}{q-s}$, $p \in P_s$. By definition, we have an edge $((S_1, S_2), (T_1, T_2))$ in $G_{C,p}$ iff $|S_1 \cap C| = \frac{q-1}{2}$ and $p \notin S_2$. The number of such S_1 is $\binom{q-s}{\frac{q-1}{2}}\binom{n-(q-s)}{\ell-\frac{q-1}{2}}$, and the number of such S_2 is $\binom{|P_s|-1}{\ell}$. Hence, Item (1) holds.

To prove Item (2), we observe that

$$z^\top A_{C,p} w = \sum_{((S_1,S_2),(T_1,T_2)) \in E(G_{C,p})} z_{S_1,S_2} w_{T_1,T_2} = \sum_{((S_1,S_2),(T_1,T_2)) \in E(G_{C,p})} \prod_{v \in S_1} x_v \prod_{p' \in S_2} y_{p'} \prod_{v \in T_1} x_v \prod_{p' \in T_2} y_{p'}$$

$$= \sum_{((S_1,S_2),(T_1,T_2)) \in E(G_{C,p})} \prod_{v \in S_1 \oplus T_1} x_v \prod_{p' \in S_2 \oplus T_2} y_{p'} = \sum_{((S_1,S_2),(T_1,T_2)) \in E(G_{C,p})} y_p \cdot \prod_{v \in C} x_v = D y_p \cdot \prod_{v \in C} x_v \,.$$

The "in particular" follows immediately from Item (2) and the definition of A and the A_i's. □

The key technical lemma, which we shall prove in Section 6.1, shows that we can find an approximately biregular subgraph of A_i for each $i \in [k]$.

Lemma 6.3 (Approximately regular submatrix)**.** *Let $A_1, \dots, A_k$ be defined as in Definition 6.1. There exists a positive integer D' with $D \geq D' \geq \frac{D}{2}$ such that the following holds. For each $i \in [k]$ and $(C, p) \in H_i$, there exists a matrix $B_{i,C,p} \in \{0,1\}^{V_L \times V_R}$ with the following properties:*

(1) $B_{i,C,p}$ is a "subgraph" of $A_{C,p}$. Namely, if $B_{i,C,p}((S_1, S_2), (T_1, T_2)) = 1$, then $A_{C,p}((S_1, S_2), (T_1, T_2)) = 1$.

(2) $B_{i,C,p}$ has exactly D' nonzero entries.

(3) The matrix $B_i \coloneqq \sum_{(C,p) \in H_i} B_{i,C,p}$ has at most $O(d_L)$ nonzero entries per row and $O(d_R)$ nonzero entries per column.

With Lemma 6.3 in hand, we can now finish the proof. Let B_i be the matrix defined in Lemma 6.3. We observe that Items (1), (2), and (3) in Lemma 6.3, along with Claim 6.2, imply that for each $x \in \{-1,1\}^n$ and $y \in \{-1,1\}^{P_s}$, there exist $z \in \{-1,1\}^{V_L}$ and $w \in \{-1,1\}^{V_R}$ such that $z^\top B w = D'\Psi(x,y)$. Hence, for any $x \in \{-1,1\}^n$ and $y \in \{-1,1\}^{P_s}$, it holds that $D'\Psi(x,y) \leq \|B\|_2 \cdot \sqrt{N_L N_R}$. We thus have that $\mathbb{E}_b[\mathrm{val}(\Psi)] \leq \frac{\sqrt{N_L N_R}}{D'} \mathbb{E}_b[\|B\|_2]$.

It remains to bound $\mathbb{E}_b[\|B\|_2]$, which we do using Matrix Khintchine (Fact 3.6) in the following claim.

Claim 6.4. Let B be the matrix defined in Lemma 6.3. Then, $\mathbb{E}_b[\|B\|_2] \leq O(\sqrt{d_L d_R k \ell \log n})$.

We postpone the proof of Claim 6.4 to the end of this section, and finish the proof of Theorem 4.3. We have that

$$\mathbb{E}_b[\mathrm{val}(\Psi)] \leq \frac{\sqrt{N_L N_R}}{D'} \mathbb{E}_b[\|B\|_2] \leq \frac{2\sqrt{N_L N_R}}{D} O(\sqrt{d_L d_R k \ell \log n})$$
$$= O(1) \cdot \sqrt{\frac{N_L d_L N_R d_R k \ell \log n}{D^2}} = \delta n O(\sqrt{k \ell \log n}),$$

as required, where we recall that $d_L = \delta n D / N_L$ and $d_R = \delta n D / N_R$.

We now finish the proof of Claim 6.4.

Proof of Claim 6.4. By Matrix Khintchine (Fact 3.6), we have $\mathbb{E}_b[\|B\|_2] \leq O(\sqrt{\sigma^2 \log(N_L + N_R)})$, where $\sigma^2 = \max(\|\sum_{i=1}^k B_i B_i^\top\|_2, \|\sum_{i=1}^k B_i^\top B_i\|_2)$. Since $B_i B_i^\top$ is symmetric, $\|B_i B_i^\top\|_2$ is bounded by the maximum ℓ_1-norm of a row in this matrix. We observe that the ℓ_1-norm of the (S_1, S_2)-th row in $B_i B_i^\top$ is simply the number of length 2 walks starting from the left vertex (S_1, S_2) in the bipartite graph with adjacency matrix B_i. As this graph has maximum left degree $O(d_L)$ and maximum right degree $O(d_R)$, it follows that this is at most $O(d_L d_R)$. Similarly, the maximum ℓ_1-norm of a row in $B_i^\top B_i$ is the number of length 2 walks starting from the right vertex (T_1, T_2) in the bipartite graph B_i, and this is at most $O(d_R d_L)$. Hence, $\|\sum_{i=1}^k B_i B_i^\top\|_2 \leq \sum_{i=1}^k O(d_L d_R) = O(k d_L d_R)$, and similarly $\|\sum_{i=1}^k B_i^\top B_i\|_2 \leq O(k d_L d_R)$ as well.

We can thus set $\sigma^2 = O(k d_L d_R)$ and apply Fact 3.6 to conclude that $\mathbb{E}_b[\|B\|_2] \leq O(\sqrt{k d_L d_R \log(N_L + N_R)})$. Recall that we have $N_L = \binom{n}{\ell}\binom{|P_s|}{\ell} \leq n^\ell (nk)^\ell \leq n^{O(\ell)}$ and $N_R = \binom{n}{\ell+1-s}\binom{|P_s|}{\ell+1} \leq n^\ell (nk)^{\ell+1} \leq n^{O(\ell)}$. Hence, $\log(N_L + N_R) = O(\ell \log n)$, which finishes the proof. ∎

6.1 Finding an approximately regular subgraph: proof of Lemma 6.3

In this section, we prove Lemma 6.3. Similar to Lemma 5.4, we will prove Lemma 6.3 by using the strategy, due to [Yan24], of bounding "conditional first moments".

Lemma 6.5 (Conditional first moment bounds). *Fix $i \in [k]$. For a left vertex (S_1, S_2), let $\deg_{i,L}(S_1, S_2)$ denote the degree of (S_1, S_2) in A_i, and for a right vertex (T_1, T_2), let $\deg_{i,R}(T_1, T_2)$ denote the right degree in A_i.*

Let $(C, p) \in H_i$, and let $\mu_{L,C,p}$ denote the distribution over left vertices that first chooses a uniformly random edge in $A_{C,p}$ and outputs its left endpoint. Similarly, let $\mu_{R,C,p}$ denote the distribution that outputs the right endpoint. Then, it holds that

$$\mathbb{E}_{(S_1,S_2)\sim\mu_{L,C,p}}[\deg_{i,L}(S_1, S_2)] \le 1 + O(1)\left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \delta n$$

$$\mathbb{E}_{(T_1,T_2)\sim\mu_{R,C,p}}[\deg_{i,R}(T_1, T_2)] \le 1 + O(1)\left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|}\delta n\,.$$

Claim 6.6 (Degree bound). Let d_L and d_R be the quantities defined in Definition 6.1. Then, for the choice of parameters given in Theorem 4.3, it holds that

$$d_L \ge \Omega\left(\left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \delta n\right) \text{ and } \left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \delta n \ge 1$$

$$d_R \ge \Omega\left(\left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|}\delta n\right) \text{ and } \left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|}\delta n \ge 1\,.$$

We postpone the proofs of Lemma 6.5 and Claim 6.6 to the end of this subsection, and now use them to finish the proof of Lemma 6.3.

Proof of Lemma 6.3 from Lemma 6.5 and Claim 6.6. Fix $i \in [k]$. Let Γ be a constant (to be chosen later), and let $V'_L = \{(S_1, S_2) : \deg_{i,L}(S_1, S_2) \le \Gamma d_L\}$, and let $V'_R = \{(T_1, T_2) : \deg_{i,R}(T_1, T_2) \le \Gamma d_R\}$.

Let $(C, p) \in H_i$. We let $A'_{i,C,p}$ be the matrix where $A'_{i,C,p}((S_1, S_2), (T_1, T_2)) = A_{C,p}((S_1, S_2), (T_1, T_2))$ if $(S_1, S_2) \in V'_L$ and $(T_1, T_2) \in V'_R$, and otherwise $A'_{i,C,p}((S_1, S_2), (T_1, T_2)) = 0$. Namely, we have "zeroed out" all rows of $A_{C,p}$ that are not in V'_L and all columns that are not in V'_R.

The left degree conditional moment bound from Lemma 6.5, combined with the lower bound on d_L from Claim 6.6 implies that $\mathbb{E}_{(S_1,S_2)\sim\mu_{L,C,p}}[\deg_{i,L}(S_1, S_2)] \le O(d_L)$. Similarly, we have $\mathbb{E}_{(T_1,T_2)\sim\mu_{R,C,p}}[\deg_{i,R}(T_1, T_2)] \le O(d_R)$. Hence, applying Markov's inequality, the number of left vertices (S_1, S_2) that are adjacent to an edge labeled by (C, p) and have $\deg_{i,L}(S_1, S_2) > \Gamma d_L$ is at most $O(D/\Gamma)$. Similarly, the number of right vertices (T_1, T_2) that are adjacent to an edge labeled by (C, p) and have $\deg_{i,R}(T_1, T_2) > \Gamma d_R$ is at most $O(D/\Gamma)$. Hence, there must be at least $D(1 - O(1/\Gamma))$ edges, i.e., nonzero entries, in $A'_{i,C,p}$.

Now, we let $B_{i,C,p}$ be any subgraph of $A'_{i,C,p}$ where $B_{i,C,p}$ has *exactly* $D' = \lfloor D(1 - O(1/\Gamma))\rfloor$ edges. This can be achieved by simply removing edges if there are too many. By choosing Γ to be a sufficiently large constant, we ensure that $D' \ge D/2$.

To prove the third property, we observe that for any left vertex (S_1, S_2), the matrix $B_i = \sum_{(C,p)\in H_i} B_{i,C,p}$ has at most $\Gamma d_L = O(d_L)$ nonzero entries in the (S_1, S_2)-th row. Indeed, this follows because it is a subgraph of the original graph A_i, and if (S_1, S_2) had degree $> \Gamma d_L$ in A_i then it has degree 0 in B_i. Similarly, any right vertex (T_1, T_2) has degree at most $O(d_R)$ in the matrix B_i. This finishes the proof. □

It remains to prove Lemma 6.5 and Claim 6.6, which we do now.

Proof of Lemma 6.5. Let $(C, p) \in H_i$ be an edge. We will first compute the left degree and then the right degree. We observe that for any $(S_1, S_2) \in V_L$, the vertex (S_1, S_2) is adjacent to at most one edge labeled by (C, p). Hence, it follows that $\mu_{L,C,p}$ is uniform over pairs of sets (S_1, S_2) such that $|S_1 \cap C| = \frac{q-1}{2}$ and $p \notin S_2$. We have

$$\mathbb{E}_{(S_1,S_2)\sim\mu_{L,C,p}}[\deg_{i,L}(S_1, S_2)] = 1 + \frac{1}{D} \sum_{(C',p')\in H_i\setminus\{(C,p)\}} |\{(S_1, S_2) : |S_1 \cap C| = |S_1 \cap C'| = \frac{q-1}{2} \text{ and } p, p' \notin S_2\}|$$

$$\leq 1 + \frac{\delta n - 1}{D} \binom{q-s}{\frac{q-1}{2}}^2 \binom{n-2(q-s)}{\ell-(q-1)} \binom{|P_s|-2}{\ell},$$

where we use that $|H_i| \leq \delta n$.

Applying Fact 3.7 and using that $|P_s| \geq 4\ell$, we have

$$\frac{1}{D} \binom{q-s}{\frac{q-1}{2}}^2 \binom{n-2(q-s)}{\ell-(q-1)} \binom{|P_s|-2}{\ell} = \frac{1}{\binom{q-s}{\frac{q-1}{2}} \binom{n-(q-s)}{\ell-\frac{q-1}{2}} \binom{|P_s|-1}{\ell}} \binom{q-s}{\frac{q-1}{2}}^2 \binom{n-2(q-s)}{\ell-(q-1)} \binom{|P_s|-2}{\ell} =$$

$$\leq O(1) \left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \frac{1}{\binom{|P_s|-1}{\ell}} \binom{|P_s|-2}{\ell} \leq O(1) \left(\frac{\ell}{n}\right)^{\frac{q-1}{2}}.$$

Hence, $\mathbb{E}_{(S_1,S_2)\sim\mu_{L,C,p}}[\deg_{i,L}(S_1, S_2)] \leq 1 + O(1) \left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \delta n$.

We now compute $\mathbb{E}_{(T_1,T_2)\sim\mu_{R,C,p}}[\deg_{i,R}(T_1, T_2)]$. As before, for any $(T_1, T_2) \in V_R$, the vertex (T_1, T_2) is adjacent to at most one edge labeled by (C, p). So, it follows that $\mu_{R,C,p}$ is uniform over pairs of sets (T_1, T_2) such that $|T_1 \cap C| = \frac{q+1}{2} - s$ and $p \in T_2$. We have

$$\mathbb{E}_{(T_1,T_2)\sim\mu_{R,C,p}}[\deg_{i,R}(T_1, T_2)] = 1 + \frac{1}{D} \sum_{(C',p')\in H_i\setminus\{(C,p)\}} |\{(T_1, T_2) : |T_1 \cap C| = |T_1 \cap C'| = \frac{q+1}{2} - s \text{ and } p, p' \in T_2\}|$$

$$\leq 1 + \frac{\delta n - 1}{D} \binom{q-s}{\frac{q+1}{2}-s}^2 \binom{n-2(q-s)}{(\ell+1-s)-(q+1)+2s} \binom{|P_s|-2}{\ell-1}$$

We have

$$\frac{1}{D} \binom{q-s}{\frac{q+1}{2}-s}^2 \binom{n-2(q-s)}{(\ell+1-s)-(q+1)+2s} \binom{|P_s|-2}{\ell-1}$$

$$= \frac{1}{\binom{q-s}{\frac{q-1}{2}}\binom{n-(q-s)}{\ell-\frac{q-1}{2}}\binom{|P_s|-1}{\ell}} \binom{q-s}{\frac{q+1}{2}-s}^2 \binom{n-2(q-s)}{(\ell+1-s)-(q+1)+2s} \binom{|P_s|-2}{\ell-1}$$

$$= \frac{1}{\binom{n-(q-s)}{\ell-\frac{q-1}{2}}\binom{|P_s|-1}{\ell}} \binom{q-s}{\frac{q+1}{2}-s} \binom{n-2(q-s)}{\ell-q+s} \binom{|P_s|-2}{\ell-1}$$

$$\leq O(1) \left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|},$$

where the last inequality is by Fact 3.7 and uses that $|P_s| \geq 4\ell$. Hence, $\mathbb{E}_{(T_1,T_2)\sim\mu_{R,C,p}}[\deg_{i,R}(T_1,T_2)] \leq 1 + O(1)\left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|}\delta n$. □

Proof of Claim 6.6. We observe that by Fact 3.7 and that $|P_s| \geq 4\ell$,

$$\frac{\delta n D}{N_L} = \delta n \cdot \frac{\binom{q-s}{\frac{q-1}{2}}\binom{n-(q-s)}{\ell-\frac{q-1}{2}}\binom{|P_s|-1}{\ell}}{\binom{n}{\ell}\binom{|P_s|}{\ell}} \geq \delta n \cdot \Omega(1)\left(\frac{\ell}{n}\right)^{\frac{q-1}{2}}.$$

Because $\ell = n^{1-2/q}\delta^{-2/q}$, we have $\left(\frac{\ell}{n}\right)^{\frac{q-1}{2}} \delta n \geq 1$, which finishes the case for the left degree.

We also have that

$$\frac{\delta n D}{N_R} = \delta n \cdot \frac{\binom{q-s}{\frac{q-1}{2}}\binom{n-(q-s)}{\ell-\frac{q-1}{2}}\binom{|P_s|-1}{\ell}}{\binom{n}{\ell+1-s}\binom{|P_s|}{\ell+1}} \geq \delta n \cdot \Omega(1)\left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|}.$$

Now, because $|P_s| \leq nk/d_s$ where $d_s = \left(\frac{\ell}{n}\right)^{s-\frac{3}{2}} k$ and $\ell = n^{1-2/q}\delta^{-2/q}$, it follows that $\delta n \left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{|P_s|} \geq \delta n \left(\frac{\ell}{n}\right)^{\frac{q+1}{2}-s} \frac{\ell}{n}\left(\frac{\ell}{n}\right)^{s-\frac{3}{2}} = \delta n \left(\frac{\ell}{n}\right)^{\frac{q}{2}} \geq 1$. This finishes the case for the right degree. □

Acknowledgements

We thank Shachar Lovett, Raghu Meka, Lisa Sauermann, and Ola Svensson for organizing a wonderful workshop at the Bernoulli Center for Fundamental Studies at EPFL on the synergies of combinatorics and theoretical computer science that led to this paper.

References

[AG24] Omar Alrabiah and Venkatesan Guruswami. Near-tight bounds for 3-query locally correctable binary linear codes via rainbow cycles. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*. IEEE, 2024.

[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.

[AGKM23] Omar Alrabiah, Venkatesan Guruswami, Pravesh K. Kothari, and Peter Manohar. A near-cubic lower bound for 3-query locally decodable codes from semirandom CSP refutation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 1438–1448. ACM, 2023.

[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.

[AOW15] Sarah R. Allen, Ryan O'Donnell, and David Witmer. How to Refute a Random CSP. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 689–708. IEEE Computer Society, 2015.

[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.

[BHKL24] Arpon Basu, Jun-Ting Hsieh, Pravesh K. Kothari, and Andrew Lin. Improved lower bounds for all odd-query locally decodable codes. 2024.

[BM16] Boaz Barak and Ankur Moitra. Noisy Tensor Completion via the Sum-of-Squares Hierarchy. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 417–445. JMLR.org, 2016.

[CGL07] Amin Coja-Oghlan, Andreas Goerdt, and André Lanka. Strong refutation heuristics for random k-SAT. *Combinatorics, Probability & Computing*, 16(1):5, 2007.

[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.

[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.

[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.

[HKM23] Jun-Ting Hsieh, Pravesh K. Kothari, and Sidhanth Mohanty. A simple and sharper proof of the hypergraph Moore bound. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pages 2324–2344. SIAM, 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.

[KM24a] Pravesh K. Kothari and Peter Manohar. An exponential lower bound for linear 3-query locally correctable codes. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 776–787. ACM, 2024.

[KM24b] Pravesh K. Kothari and Peter Manohar. Exponential lower bounds for smooth 3-lccs and sharp bounds for designs. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*. IEEE, 2024.

[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.

[LP91] Françoise Lust-Piquard and Gilles Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

[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.

[RRS17] Prasad Raghavendra, Satish Rao, and Tselil Schramm. Strongly refuting random CSPs below the spectral threshold. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 121–131. ACM, 2017.

[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.

[Yan24] Tal Yankovitz. A stronger bound for linear 3-lcc. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*. IEEE, 2024.

[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.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il