THEORY OF COMPUTING, Volume 13 (11), 2017, pp. 1–36
www.theoryofcomputing.org

Superquadratic Lower Bound for 3-Query Locally Correctable Codes over the Reals

Zeev Dvir* Shubhangi Saraf† Avi Wigderson‡

Received July 20, 2015; Revised March 20, 2017; Published October 23, 2017

Abstract: We prove that 3-query linear locally correctable codes of dimension d over the reals require block length $n > d^{2+\alpha}$ for some fixed, positive $\alpha > 0$. Geometrically, this means that if n vectors in $\mathbb{R}^d$ are such that each vector is spanned by a linear number of disjoint triples of others, then it must be that $n > d^{2+\alpha}$. This improves the known quadratic lower bounds (e. g., Kerenidis–de Wolf (2004), Woodruff (2007)). While the improvement is modest, we expect that the new techniques introduced in this article will be useful for further progress on lower bounds of locally correctable and decodable codes with more than 2 queries, possibly over other fields as well.

Several of the new ideas in the proof work over every field. At a high level, our proof has two parts, *clustering* and *random restriction*.

The clustering step uses a powerful theorem of Barthe from convex geometry. It can be used (after preprocessing our LCC to be *balanced*), to apply a basis change (and rescaling) of the vectors, so that the resulting unit vectors become *nearly isotropic*. This together with the fact that any LCC must have many "correlated" pairs of points, lets us deduce that the

An extended abstract of this paper appeared in the Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing 2014 [14].

*Supported by NSF CAREER award DMS-1451191 and NSF grant CCF-1523816.

†Supported by NSF grant CCF-1350572.

‡Supported by NSF grant CCF-1412958.

ACM Classification: E.4

AMS Classification: 94B65, 52C35

Key words and phrases: coding theory, discrete geometry, combinatorics, lower bounds

© 2017 Zeev Dvir, Shubhangi Saraf, and Avi Wigderson
Licensed under a Creative Commons Attribution License (CC-BY)
DOI: 10.4086/toc.2017.v013a011

vectors must have a surprisingly strong geometric clustering, and hence also combinatorial clustering with respect to the spanning triples.

In the restriction step, we devise a new variant of the dimension reduction technique used in previous lower bounds, which is able to take advantage of the combinatorial clustering structure above. The analysis of our random projection method reduces to a simple (weakly) random graph process, and works over any field.

1 Introduction

Locally-correctable codes (sometimes under different names of program self-correctors or random self-reductions), abbreviated LCCs, have the property that each symbol of a corrupted codeword can be recovered, with high probability, by randomly accessing only a few other symbols. LCCs have played a key role in important developments within several (impressively) diverse areas of theoretical computer science, which we briefly summarize below.

Blum and Kannan [9] introduced the idea of probabilistic, local correction for the purpose of program checking. With the follow-up papers [10] on linearity testing and [27] on low-degree testing this sequence inaugurated the field of Property Testing and Sublinear Algorithms. The realization of [25, 7], that Reed-Muller codes (namely low-degree multivariate polynomials) are locally correctable, gave the first random self-reducibility examples of very hard functions like the Permament, and this average-case to worst-case complexity reduction was useful for pseudo-random generators [4]. It further lead (with many more ideas) to the celebrated sequence of characterizations of the power of probabilistic proofs, IP = PSPACE by [26, 28], MIP = PSPACE by [3] and PCP = NP by [2, 1]. Close cousins of LCCs, Locally-*Decodable* Codes (LDCs),[1] formally introduced in [19] but having their origins in these earlier works, were key to Private Information Retrieval and other models of secure delegation of computation (see, e. g., [11]). Dvir [12] has shown that sufficiently strong lower bounds on LCCs would yield explicit rigid matrices, which are related, via the work of Valiant [29] to circuit complexity.[2] While this has not materialized yet, it motivated the invention of *multiplicity codes* by [23] which are new LCCs of high rate, and turn out to yield optimal list-decodable codes as well [22]. Finally, since the work of Dvir and Shpilka [16], LDCs and LCCs have played a role in understanding basic problems in Polynomial Identity Testing and established its connection to problems in Incidence Geometry, e. g., [20, 5, 15].

The most important parameters of LCCs are the number of queries, q, made by the correcting algorithm, and the block length n as a function of the message length (or dimension, for linear codes) d, where we fix corruptions to some small fixed fraction, say 1%. For upper bounds, the best constructions we have are still based on Reed-Muller codes[3] which exist only over finite fields. For q queries these require block length about $\exp(d^{1/(q-1)})$. Indeed most applications require the block-length n to be polynomial in d and hence using these codes forces the number of queries to be at least logarithmic.

[1] In LDCs one needs to locally recover *only* d linearly independent coordinates (equivalently, the message) from the corrupted codeword, rather than all n of them.

[2] While the work of Kopparty, Saraf, and Yekhanin shows that, over small finite fields, this approach could not give superlinear circuit lower bounds, the approach might still be valid over large fields.

[3] For the weaker LDCs there are far better constructions, based on the work of Yekhanin and Efremenko [32, 17, 13], but these are not known to be locally correctable.

Finding better codes, and in particular constant-query, polynomial block-length LCCs, has been a major challenge, and this challenge naturally turns attention to the limits of constant query LCCs and LDCs.

On the lower bound front, relatively little is known to rule out the feasibility of the challenge above. We shall restrict ourselves to *linear* codes[4] over a field $\mathbb{F}$, namely when the set of codewords is a subspace of $\mathbb{F}^n$ of dimension d. We will denote by q-LCC such a linear locally-correctable codes with q queries. It is easy to see that 1-LCCs do not exist over any field. The first set of interesting results came for 2-LCCs, and here strong lower bounds are known through a variety of techniques. An exponential $n > 2^{\Omega(d)}$ lower bound via isoperimetric/entropy methods for 2-LCCs over $\mathbb{F}_2$ follows from the same methods as for the (weaker) LDCs [18, 21, 16] and is matched by the Hadamard code whose generating matrix is composed of all binary vectors over $\mathbb{F}_2$. Strangely, while these vectors provide an LDC over *every* field, they fail to be an LCC except over $\mathbb{F}_2$. This gap was first explained in [5, 15] where the authors showed that over the real numbers (and indeed even over large enough finite fields), 2-LCCs simply do not exist! For every error-rate δ the dimension d for which such codes exist is finite, and cannot exceed $\mathsf{poly}(1/\delta)$. The proofs here use a combination of geometric, analytic and linear-algebraic techniques, and give quantitative form to known qualitative point-line incidence theorems. Tighter bounds of $n > p^{\Omega(d)}$ over finite fields of prime order p were proved in [8] using methods from arithmetic combinatorics, matching the trivial construction of taking all vectors in $(\mathbb{F}_q)^d$.

For $q \geq 3$ the known lower bounds are far weaker, and practically only one lower-bound technique is known: random restrictions of the given code which reduce the number of queries q to 2 or 1, appealing to the lower bounds above. This technique was introduced for LDCs by Katz and Trevisan [19]. The resulting lower bounds trivially hold for (the stronger) LCCs as well. The best bounds known are due to [21, 30], which show that linear q-LDCs, over any field $\mathbb{F}$, must satisfy $n = \widetilde{\Omega}(d^{1+1/(\lceil q/2 \rceil - 1)})$ for every $q \geq 3$. So, in particular, the best lower bound for 3-LDCs (or LCCs) is quadratic, $n = \widetilde{\Omega}(d^2)$. (For linear codes the $\widetilde{\Omega}$ was replaced by Ω in [31].) Our main result is breaking this quadratic barrier for 3-LCCs over the real numbers. Over the reals there are no known constructions of constant-query LCCs (of any rate!). We prove that for some fixed constant[5] $\alpha > 0$ every linear 3-LCC over the reals must satisfy $n = \Omega(n^{2+\alpha})$, even when the error parameter δ is allowed to be polynomially small in n. To this end, we introduce several new ideas and techniques, which we hope will lead to further progress. Some of our ideas are general enough to work over any field, while others are specifically tailored for the reals. We briefly discuss now the main sources for our improvement over the known quadratic lower bound. A more detailed overview of the proof is given after the formal statement of the theorem in the next section.

1.1 Clustering and restrictions

A linear 3-LCC of dimension d and block length n over $\mathbb{F}$ may be viewed as a set $V \subset \mathbb{F}^d$ of n vectors (which form the generating matrix of the code), together with n collections M_v, one for each $v \in V$. Each M_v is a matching of δn disjoint triples from V, and each of the triples in M_v spans v. This structure is easy to deduce for linear codes from the more traditional definition using a randomized decoder (cf. Definition 2.1).

We now informally describe a way to obtain a possible quadratic lower bound on n, which uses

[4] Some of the results below are known also for non-linear codes.

[5] We did not make an attempt to optimize the constant α, but the proof gives some $\alpha > .01$.

random restriction to reduce the dimension of the code. Pick a set $A \subset V$ of size about $\sqrt{n}$ at random. Then, take a linear projection whose kernel is exactly the span of the vectors in A and apply it to the elements of V. Notice that in expectation, for every $v \in V$, a pair of points in A will be contained in some triple in M_v. Thus, after the projection the third point in that triple will become the same as v (up to scaling). As this happens to every point, we expect V to shrink by a factor of 2! Notice that in such a projection, the dimension of V can drop by at most $|A| \approx \sqrt{n}$. Repeating this process logarithmically many times will shrink V completely, revealing that its original dimension could not have been larger than $\sqrt{n}\log n$, giving a near quadratic relation $n \geq d^2/\log d$. We note that the proofs appearing in the literature are somewhat different then the one we just described. Indeed, there are several possible ways of using a random restriction argument to get a quadratic bound (up to poly-logarithmic factors) for linear 3-LCCs. The argument above is new to this paper, and is indeed a simplified variant of our actual proof, which improves its analysis over the reals.

It is not hard to see that if the collection of triples in all of matchings M_v were chosen at random, the analysis above could not be improved. But a random collection is far from being an LCC. Indeed, in contrast to standard codes, which exist in abundance and a random subspace is one with high probability, locally correctable (or decodable, or testable) codes are extremely rare and structured. This raises the question of what other structural properties are imposed on the matchings M_v in an LCC. In this paper we reveal a new such property, *clustering*, at least when the underlying field is the reals.[6] We conclude with a simplified description of this clustering property, how it is obtained, and how it enables better analysis of the random restriction process.

A collection $\{M_v\}$ of matchings of triples is said to be *clustered* if there are about $\sqrt{n}$ subsets $S_1, \ldots, S_{\sqrt{n}}$ of V, each of size about $\sqrt{n}$, such that *every* triple in *every* matching M_v has a pair in one of these sets. Note that such a configuration is extremely far from random. Indeed, as these sets have at most $n^{3/2}$ pairs between them, many of the triples (of different matchings) share pairs (a typical pair exists in about $\sqrt{n}$ triples!). Note that this cluster structure is completely combinatorially described.

Why should the triples in a 3-LCC admit such a clustering? The main observation is that, over the reals, a small linearly dependent subset, such as a 4-tuple composed of v and a triple from M_v, must contain a pair which is significantly correlated (say, with inner product at least $1/4$ for said example). Thus, a 3-LCC must contain many correlated pairs. On the other hand, a powerful result of Barthe from convex geometry allows us to deduce that, after a carefully chosen change of basis, the vectors of our code are almost isotropic, meaning that they point roughly equally in all directions in space. This implies that most pairs are hardly correlated at all. These two seemingly contradicting structures can exist only if the points in V are *geometrically* clustered. Delicate analysis shows that they can be partitioned into roughly $\sqrt{n}$ small balls. The correlations then must arise from triples containing a pair in one of the (geometric) clusters.

Why does clustering help? Let us return to the random restriction and projection argument above, but let us pick now the set A as follows. First pick one of the clusters S_i uniformly at random, and inside it pick A at random of size about $n^{1/4}$. The clustering ensures that this much smaller set has a pair intersecting each of the matchings M_v in expectation (due to the fact that a typical pair in a typical cluster participates in $\sqrt{n}$ matchings). So a much smaller set A suffices to create the same effect after projection, namely a shrinking of the set V by a factor of 2. Again a logarithmic number of such restrictions is

[6] The actual proof requires several extra conditions on the code, which can be obtained via a sequence of reductions.

likely to shrink V completely, giving a dimension upper bound of $n^{1/4}\log n$, and yielding the lower bound $n \geq d^4/\log d$. We note again that this part works over any field, as long as the triples are clustered.

"Balanced" codes. A recurring notion in our proof is that of an LCC in which no large subset of the coordinates lies in a subspace of significantly lower dimension. One can think of such codes as being "balanced" in the sense that they cannot be "compressed" (by projecting the large set of low dimension to zero). Our proof contains a sequence of reductions, used to obtain certain conditions that are used in the clustering and restriction steps. Each of these reductions can only be carried out if the code is "balanced" and this property is used in several different ways in the proof. If the code is not "balanced" we can use an iterative argument that projects the large low-dimensional subset to zero. We find this condition of being balanced a very natural one in the context of LCCs (and other codes) and hope it could be useful as a conceptual tool in future works.

Organization. In Section 2 we state our results formally. Then, in Section 3 we provide a more detailed and technical overview of the proof. The organization of the rest of the paper (which contains a complete proof of our main result) is given at the end of Section 3.

Acknowledgments. We thank the anonymous referees for their careful reading of the paper and for many useful comments. We are grateful to Boaz Barak, Moritz Hardt and Amir Shpilka for their contribution in early stages of this work. In particular, we thank Moritz Hardt for introducing us to Barthe's work.

2 Definitions and results

For a string $y \in \mathbb{F}^n$, we define $w(y)$ to be the number of nonzero entries in y. A q-matching M in $[n]$ is defined to be a set of disjoint unordered q-tuples (i. e., disjoint subsets of size q) of $[n]$.

Definition 2.1 (Linear q-LCC, decoder definition)**.** A linear (q,δ)-LCC of dimension d over a field $\mathbb{F}$ is a d-dimensional linear subspace $U \subset \mathbb{F}^n$ such that there exists a randomized decoding procedure $D : \mathbb{F}^n \times [n] \to \mathbb{F}$ with the following properties.

1. For all $x \in U$, for all $i \in [n]$ and for all $y \in \mathbb{F}^n$ with $w(y) \leq \delta n$ we have that $D(x+y,i) = x_i$ with probability at least $3/4$ (the probability is taken only over the internal randomness of D).

2. For every $y \in \mathbb{F}^n$ and $i \in [n]$, the decoder $D(y,i)$ reads at most q positions in y.

Definition 2.2 (Linear q-LCC, geometric definition)**.** Let $V = (v_1,\ldots,v_n) \in (\mathbb{F}^d)^n$ be a list of n vectors spanning $\mathbb{F}^d$. We say that V is a linear (q,δ)-*LCC* in geometric form if for every $v \in V$ there exists a q-matching M_v in $[n]$ of size at least δn such that for every q-tuple $\{j_1,\ldots,j_q\} \in M_v$ it holds that $v \in \mathsf{span}\{v_{j_1},\ldots,v_{j_q}\}$.

It is well known that any linear (q,δ)-LCC (over any field) can be converted into the geometric form given above by replacing δ with δ/q. The transformation is simple. take $v_1,\ldots,v_n \in \mathbb{F}^d$ to be the rows of

the generating matrix of U. Clearly, this does not change the dimension of the code. This is surprising since it implies also that the decoder in the first definition can be made non adaptive without much loss in parameters (for linear codes).

In our results we will assume that the error parameter δ is not too small as a function of n. Specifically, we will require that $n \geq (1/\delta)^{\omega(1)}$. This condition can be replaced with $n \geq (1/\delta)^C$ for a sufficiently large absolute constant C which can be calculated from the proof.

We now state our main result which bounds the dimension of 3 query LCC's when the underlying field is $\mathbb{R}$.

Theorem 2.3 (Main Theorem)**.** *There exists an absolute constant $\varepsilon > 0$ such that if $V = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$ is a linear $(3, \delta)$-LCC and $n \geq (1/\delta)^{\omega(1)}$, then*

$$d = \dim(V) \leq n^{1/2-\varepsilon}.$$

3 Proof overview: "Cluster and Restrict" method

From a high level, our proof is divided into two conceptually distinct steps.

1. *Clustering step.* Show that the triples used in the matchings $M_v, v \in V$ are "clustered" in some precise sense (described below).

2. *Restriction step.* Use the clustering to find a large subset of V that has low dimension. The name of this step is due to the fact that it uses a random restriction argument (projecting a random subset to zero).

Combining these two steps (in Lemma 10.1) we get that V must have a large subset (of size $\Omega(n)$) with low dimension (at most $n^{1/2-\varepsilon}$). Using this to prove a global dimension bound on V (as in Theorem 2.3) is done using a standard amplification lemma (Lemma 10.2) similar to that in [5, 8]. For simplicity, we will use big "O" notation to hide constants depending on δ (only for this overview).

We now describe each of these steps in more detail. The fact that V is a code over $\mathbb{R}$ is only used in the clustering step. The restriction step works over any field, provided that the triples are already clustered. A recurring theme in the proof is that we are always free to assume that V does not have a large subset of low dimension. Another recurring operation is "sending a subset U of V to zero." By this statement we mean: pick a linear map A whose kernel is $\mathsf{span}(U)$ and apply it to all the elements of V. We will use the simple fact that, if $\dim(U) = r$ and $\dim(V) = d$ then $\dim(A(V))$ is at least $d - r$, where $A(V)$ is the list of vectors $A(v), v \in V$.

The clustering step is given by Lemma 8.2 which we state now in an informal form. We will elaborate below on the two conditions appearing in the lemma (the well-spread vectors condition and the low triple-multiplicity condition). Recall that V is associated with n 3-matchings $M_v, v \in V$ used in the decoding.

Lemma 3.1 (Informal statement of Lemma 8.2)**.** *Suppose V is a $(3, \delta)$-LCC that satisfies the well-spread vectors condition and the low triple-multiplicity condition and suppose that $d > n^{1/2-\varepsilon}$. Then there are subsets $S_1, \ldots, S_m \subset V$ (not necessarily disjoint) so that*

1. *for each* $i \in [m]$, $|S_i| \leq O(n^{1/2+\varepsilon})$;
2. $\Omega(n^{1/2-\varepsilon}) \leq m \leq O(n^{1/2+\varepsilon})$;
3. *each triple in each matching* M_v *has two of its elements in one of the sets* S_i.

Before we explain the two conditions in the lemma, i. e., the well-spread vectors condition and the low triple-multiplicity condition, notice that the existence of sets $S_1, \ldots, S_m$ as above is something that does not hold for a "typical" family of $\Omega(n^2)$ triples. In fact, if the triples were chosen at random there would not be such sets with probability close to one. Referring to the sets S_i as "clusters" is also justified by the fact that they actually form clusters in $\mathbb{R}^d$ (i. e., they are all correlated with some fixed point). This geometric fact, however, is not used anywhere in the proof—all we need is the combinatorial structure. We now explain the two conditions on the code V mentioned in the lemma.

- **Well-spread vectors condition.** The vectors $v_1, \ldots, v_n$ comprising V should be is some sense well spread. Observe that w. l. o. g. by a suitable scaling to each vector, we can assume that the vectors $v_1, \ldots, v_n$ are unit vectors, and we will make this assumption. Formally, we require that for every unit vector $w \in \mathbb{R}^d$ we have $\sum_{i \in [n]} \langle v_i, w \rangle^2 \leq O(n^{1/2+\varepsilon})$. This means, in particular, that every small ball can contain at most $O(n^{1/2+\varepsilon})$ vectors. Clearly, a general LCC V does not need to satisfy this condition. For example, if V has a large subset of low rank such a statement cannot hold (using a pigeon hole argument on the unit sphere in low dimension). We are able, however, to reduce to this case using Lemma 6.1, which uses a powerful result of Barthe (Lemma 5.1) that is developed in Section 5. Roughly speaking, Barthe's theorem can be used to show that, unless V has a large subset of low rank there is an invertible linear map M on $\mathbb{R}^d$ so that, if we replace each v_i with $Mv_i/\|Mv_i\|$, the well-spread vectors condition is satisfied. The proof of this result (part of which appear in Section 5) uses tools from convex geometry. We derive a particularly convenient form of Barthe's theorem as Theorem 6.4 which might be of independent interest.

- **Low triple-multiplicity condition.** This condition requires that a single triple does not appear in "too many" (roughly $n^{O(\varepsilon)}$) different matchings. In Section 7 we prove Lemma 7.2 which shows how to reduce to this case, assuming V does not have a large subset of low rank. The reduction uses the fact that if a single triple is used in too many matchings, then projecting the elements in this triple to zero causes many other points to go to zero. If a point v is mapped to zero as a result, and if v is used in many triples (say $\Omega(n)$) all of these triples "become" pairs when v maps to zero. Using this observation, we show that we can send a relatively small number of points to zero and construct a 2-query locally decodable code (LDC) of relatively high dimension. We then apply the known bounds for 2-query LDCs (these are variants of LCCs and described in Section 4) to get a contradiction. This reduction is also field independent and does not use any properties of the real numbers.

The main observation leading to clustering is that we can assume, w. l. o. g., that all triples $(i, j, k) \in M_v$ are so that the three vectors v_i, v_j, v_k are almost orthogonal to v. This follows directly from the well-spread vectors condition by bounding from above the number of vectors correlating with v and discarding the corresponding triples from M_v (for each $v \in V$). Once we have this condition, we observe that since v, v_i, v_j, v_k are linearly dependent and, since v is not correlated with the other three vectors, we must have

that v_i, v_j, v_k are close to being in a two dimensional plane. (Recall that these are all unit vectors.) This means that in each triple there must be two elements that are correlated with each other! This is already a non trivial fact, in particular since we know (by the well-spread vectors condition) that each point cannot be correlated with many other points.

Proceeding with a more careful analysis of the different types of triples that can arise, and using some graph theoretic arguments, we arrive at the required clusters. In this step we use the bound on the maximum triple-multiplicity.

Note that the clustering lemma implies that there are many pairs in $V \times V$ that appear in many triples. This is due to the simple upper bound of $n^{1.5+O(\varepsilon)}$ on the total number of possible pairs in all of the clusters $S_1, \ldots, S_m$ and the fact that together they cover pairs from a quadratic number of triples. This should be contrasted with the results of [5, 15] which prove strong lower bounds for q-LCC's (for any constant q) in which every pair is in a bounded number of triples (these are called "design" LCCs).

3.1 Restriction step

The restriction step (given in Lemma 9.1) shows that if V satisfies the clustering condition (given in Lemma 8.2) then it contains a large subset of low rank. We now state a simplified form of this lemma.[7]

Lemma 3.2 (Informal statement of Lemma 9.1)**.** *Let $\mathbb{F}$ be a field. Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ be a $(3, \delta)$-LCC with matchings $M_v, v \in V$. Suppose there exist sets $S_1, \ldots, S_m \subset [n]$ as in Lemma 8.2, clustering the triples in the matchings M_v. Then, there is a subset $V' \subset V$ of size $|V'| \geq (\delta/2)n$ and dimension at most $n^{1/2-\varepsilon}$.*

This step is called the "restriction step" since it uses the "clusters" $S_1, \ldots, S_m$ found in the clustering step to show (Lemma 9.2) that there is a small set $U \subset V$ (of size roughly $n^{1/4+7\varepsilon}$) such that, projecting all elements of U to zero, reduces the dimension of V to at most $n^{10\varepsilon}$. This will imply a dimension bound of $n^{1/4+7\varepsilon} + n^{10\varepsilon}$ on the initial dimension of V. (The reason we do not get a $n^{1/4+7\varepsilon}$ upper bound on the dimension of V is due to the clustering step.)

The starting point for the proof of this lemma is the following simple observation. If v is spanned by a triple (v_i, v_j, v_k), then projecting two elements of that triple, say v_i, v_j, to zero makes the two vectors v, v_k proportional to each other. (This uses the fact that v is not spanned by any proper subset of the triple, and we can easily reduce to this case.) Now, suppose that there are t triples in the code that have at least two element in U. Then projecting U to zero makes makes t pairs of vectors proportional to each other (as in the v, v_k example). Consider the graph on vertex set V in which we add an edge for each proportional pair v, v_k obtained by sending a pair $v_i, v_j \in U$ in a triple $(v_i, v_j, v_k) \in M_v$ to zero. Since the property of being proportional to each other is an equivalence relation on $\mathbb{R}^d$, we can bound the dimension of V after projecting U to zero by the number of connected components of the graph.

This leaves us with the task of finding a set U so that the resulting graph has at most $n^{10\varepsilon}$ components. To find such a U we use a probabilistic argument. We will pick U at random according to a particular distribution and then argue that the expected number of connected components is small. To pick the random U we proceed in $r \sim n^{4\varepsilon}$ steps as follows. In each step pick one of the clusters S_i at random and

[7] One should not expect things to get better the larger ε is (as this lemma might suggest) since the condition $d > n^{1/2-\varepsilon}$ appearing in the clustering lemma prevents ε from being too large.

then pick a random subset of S_i of size $\sim n^{1/4+3\varepsilon}$. The union of these sets will be U. The upper bound on the expected number of components is derived by considering the (expected) reduction in the number of connected components in each of the r steps. Consider some connected component and let v be some vector in it. We can assume the component is not too large, since the number of large components is trivially bounded (large being close to $n^{1-\varepsilon}$). Since each M_v is a matching, the random choice of the vectors in the i'th step will (with good probability) add an edge to v with a neighbor that is not likely to land in the connected component containing v. Hence, with good probability the connected component will "merge" with another component. Carefully analyzing this process gives us the required bound.

3.2 Organization

We begin with some general preliminaries and notation in Section 4. In Section 5 we describe (and sketch the proof of) Barthe's theorem which is used in Section 6 to reduce to the case that the points in V are well-spread. In Section 7 we show how to reduce to the case that V has low triple multiplicities. Section 8 contains the proof of the clustering step and Section 9 contains the proof of the restriction step. Finally, in Section 10 we show how to put all the ingredients together and prove Theorem 2.3.

4 General preliminaries

4.1 Choice of notation

Lists vs. multisets. The reason we are treating V as a list and not as a set is that V might have repetitions. For instance u and v might be distinct elements in the list V, but might correspond to the same vector in $\mathbb{F}^d$. The repetition corresponds to the fact that there might be repeated columns in the generator matrix of the code, which may potentially make the property of local correction easier to satisfy. Indeed in the recent lower bounds for 2-query LCCs [8, 5], handling the fact that there might be repetitions added significant complexity to the proofs of the lower bounds. In the current paper too we deal with repetitions by treating V as a list. An equivalent treatment would be to treat V as a multiset, and we make no distinction between these notions. We think of a multiset as an ordered list of elements which might contain repeated elements. If A is a multiset/list, we call B a subset of A if B is another multiset/list obtained by taking a subset of A. We will say that B and C are *disjoint* subsets of A if they are both obtained from sub-lists on disjoint subsets of the indices. When referring to the *size* of a multiset we will always count the number of elements *with* multiplicities (unless we state explicitly that we are counting *distinct* elements).

Although we defined a matching to be a set of tuples in $[n]$, when we are dealing with a specific list $V = (v_1, \ldots, v_n)$, we might identify a tuple $(j_1, \ldots, j_q)$ of a matching with the tuple $(v_{j_1}, \ldots, v_{j_q})$, and we use these two notions interchangeably. Moreover, a matching M_v denotes the matching corresponding to a particular element $v \in V$, and if u and v are different elements of V, even if they correspond to the same vector in $\mathbb{F}^d$, then M_u and M_v could be different matchings.

4.2 Basic operations on LCCs

For a list $V \in (\mathbb{R}^d)^n$ we denote by $\mathsf{span}(V)$ the subspace spanned by elements of V and by $\dim(V)$ the dimension of this span.

The following simple observation shows that a sufficiently large subset of an LCC is also an LCC.

Claim 4.1. *If $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ is a $(3, \delta)$-LCC and $U \subset V$ is of size $|U| \geq (1 - \delta/2)n$ then U is a $(3, \delta/2)$-LCC of the same dimension as V. Moreover, if $M_v, v \in V$ are any matchings used in the decoding of V then we can take the matchings for the new code U to be subsets of the old matchings.*

Proof. Observe that in each matching M_v, there are at most $(\delta/2)n$ triples that contain an element outside U. Thus, in U we could construct matchings of size $(\delta/2)n \geq (\delta/2)|U|$. The claim about the dimension follows from the fact that U contains triples spanning all of the elements of V (not just those in U). □

Another simple observation is that applying an invertible linear map to the elements of V preserves the property of being an LCC.

Observation 4.2. *If $V = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$ is a $(3, \delta)$-LCC then, for any invertible linear map $M : \mathbb{R}^d \to \mathbb{R}^d$ the list $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_n) \in (\mathbb{R}^d)^n$, with $\hat{v}_j = Mv_j/\|Mv_j\|$, is also a $(3, \delta)$-LCC.*

4.3 Lower bounds for 2-query LDCs

One of the ingredients in the proof will be a strong (exponential) lower bound on the length of linear 2-query Locally Decodable Codes (LDCs), which are weaker versions of LCCs. As with LCCs there are two ways of defining LDCs.

Definition 4.3 (linear q-LDC, decoder definition)**.** A linear (q, δ)-LDC over a field $\mathbb{F}$ is a linear d-dimensional subspace $U \subset \mathbb{F}^n$, and a set of d coordinates $j_1, j_2, \ldots j_d \in [n]$ such that the projection of U on to those d coordinates is full dimensional,[8] and such that there exists a randomized decoding procedure $D : \mathbb{F}^n \times [d] \to \mathbb{F}$ with the following properties:

1. For all $x \in U$, for all $i \in [d]$ and for all $y \in \mathbb{F}^n$ with $w(y) \leq \delta n$ we have that $D(x+y, i) = x_{j_i}$ with probability at least $3/4$. (The probability is taken only over the internal randomness of D.)
2. For every $y \in \mathbb{F}^n$ and $i \in [d]$, the decoder $D(y, i)$ reads at most q positions in y.

Let $\{e_1, e_2, \ldots, e_d\}$ be the set of standard basis vectors in $\mathbb{R}^d$.

As with LCCs, taking the rows of the generating matrix (and possibly applying an invertible linear map that sends them to the e_is) allows us to move to the geometric form. This might require us to replace δ with δ/q.

Definition 4.4 (linear q-LDC, geometric definition)**.** Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ be a list of n vectors spanning $\mathbb{F}^d$. We say that V is a linear (q, δ)-*LDC* in geometric form if for every $i \in [d]$ there exists a q-matching M_i in $[n]$ of size at least δn such that for every q-tuple $\{v_{j_1}, v_{j_2}, \ldots, v_{j_q}\} \in M_i$ it holds that $e_i \in \mathsf{span}\{v_{j_1}, v_{j_2}, \ldots, v_{j_q}\}$. We denote by $d = \dim(V)$.

[8] If the LDC was systematic, then the first d coordinates would suffice.

Theorem 4.5 (lower bounds for 2-LDC [16]). *Let $\delta \in [0,1]$, $\mathbb{F}$ be a field, and let $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ be a linear $(2,\delta)$-LDC in geometric form. Then*

$$n \geq 2^{\frac{\delta d}{16}-1}.$$

4.4 Codes in regular form

In the restriction step, it is convenient for us to assume that for each triple $(v_i, v_j, v_k) \in M_v$ each element of the triple is "used" in decoding to v. Indeed in Claim 4.7, we show how we can easily reduce to this case provided that no large subset of V has low rank. More precisely, for $x, y, z \in \mathbb{R}^d$, let us denote by $\mathsf{span}^*\{x,y,z\}$ the set of all elements of the form $\alpha x + \beta y + \gamma z$ with $\alpha, \beta, \gamma \in \mathbb{R}$, such that α, β, γ are all nonzero.

Definition 4.6. Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ be a $(3,\delta)$-LCC with decoding matchings $M_v, v \in V$. We say that V (with these matchings) is in *regular* form if, in each triple $(x,y,z) \in M_v$ we have that $v \in \mathsf{span}^*\{x,y,z\}$.

Claim 4.7. *Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ be a $(3,\delta)$-LCC so that every subset $U \subset V$ of size $|U| \geq (\delta/2)n$ has dimension at least $\omega((1/\delta)\log(n))$. Then, there exists a $(3,\delta/4)$-LCC $V' \subset V$ of size $n' \geq (1-\delta/2)n$, and dimension $d' = d$, that is in regular form. Moreover, given any matchings M_v for the code V we can take the new (regular) matchings M'_v for V' to be sub-matchings of the original ones.*

Proof. Call a triple $(x,y,z) \in M_v$ *bad* if there is a proper subset of it that spans v, i. e., $v \notin \mathsf{span}^*\{x,y,z\}$. If there were $(\delta/2)n$ points $v \in V$, each with at least $(\delta/10)n$ bad triples in M_v, then we could use these bad triples to construct a $(2,\delta/10)$-LDC of size less than n decoding $\omega((1/\delta)\log(n))$ linearly independent elements of V. This would give a contradiction using Theorem 4.5 and the assumption on the dimension of any set of size $(\delta/2)n$ in V. Therefore, there are at most $(\delta/2)n$ points $v \in V$ with many (at least $(\delta/10)n$) bad triples. Throwing away this set, and removing all triples containing them (as well as all bad triples from the other matchings) gives us the code V' a required (as in Claim 4.1). □

5 Barthe's theorem

The main purpose of this section is to derive Lemma 5.1, a result of F. Barthe [6] which, given a set of points sufficiently close to being in general position, finds a linear transformation that "moves" these points so that their "directions" point in a close to uniform way. More precisely, for a set $U = (u_1, \ldots, u_n) \in (\mathbb{R}^d)^n$ let $\mathcal{B}(U)$ be the set of all subsets of $[n]$ of size d such that the corresponding vectors of U form a basis of $\mathbb{R}^d$. Suppose that there is a distribution μ supported on $\mathcal{B}(U)$ such that when sampling a random basis from μ, each element of U is chosen with some good probability. Then there is an invertible linear transformation such that after normalizing, the new points are "approximately isotropic." This result is formalized in Lemma 5.1 which we state below.

Lemma 5.1. *Let $U = (u_1, \ldots, u_n) \in (\mathbb{R}^d)^n$. Let $S \subseteq [n]$, and suppose μ is a distribution supported on $\mathcal{B}(U)$ such that for all $j \in S$*

$$\alpha \leq \Pr_{I \sim \mu}[j \in I].$$

Then, there exists an invertible linear map $M : \mathbb{R}^d \to \mathbb{R}^d$ so that, denoting $\hat{u}_j = Mu_j/\|Mu_j\|$, we have for all unit vectors $w \in \mathbb{R}^d$

$$\sum_{j\in S} \langle \hat{u}_j, w\rangle^2 \leq \frac{2}{\alpha}.$$

Observe that if the vectors are in general position then the uniform distribution on distinct d-tuples gives $\alpha = d/n$, in which case we would get

$$\sum_{j\in[n]} \langle \hat{u}_j, w\rangle^2 \leq \frac{2n}{d}.$$

One can just assume the lemma above which follows in a straightforward way from from [6], and skip to the next section. However for completeness, we present a proof here. Before we give the proof, we first set up some notation.

For a finite set S, a distribution supported on S is a function $\mu : S \to [0,1]$ so that $\sum_{x\in S} \mu(x) = 1$. For two vectors $u, v \in \mathbb{R}^d$ we denote by $u \otimes v$ the tensor product of u and v, namely the $d \times d$ matrix with entries $A_{ij} = u_i v_j$. We denote by $I_{d\times d}$ the $d \times d$ identity matrix. For $u \in \mathbb{R}^d$ we denote by $\|u\|$ the Euclidean (or ℓ_2) norm.

Definition 5.2 ($\mathcal{B}(U), \mathcal{K}(U)$)**.** Let $U = (u_1, \ldots, u_n) \in (\mathbb{R}^d)^n$ be a list of n points. Let $I \subseteq [n]$. We denote by $U_I = (u_i)_{i\in I}$ the sub-list of U with indices in I. We denote by

$$\mathcal{B}(U) = \{I \subset [n] \mid U_I \text{ is a basis of } \mathbb{R}^d\}$$

the set of index sets corresponding to sub-lists of U of length d which are linearly independent (and so span $\mathbb{R}^d$). For each $I \subset [n]$ we let $1_I \in \mathbb{R}^n$ denote the indicator vector of the set I. Finally we denote by $\mathcal{K}(U) \subset \mathbb{R}^n$ the convex hull of the vectors 1_I for all $I \in \mathcal{B}(U)$. We denote by $\mathcal{K}(U)^o$ the relative interior of $\mathcal{K}(U)$.[9]

Claim 5.3 (Properties of $\mathcal{K}(U)$)**.** *Let $U = (u_1, \ldots, u_n) \in (\mathbb{R}^d)^n$ be a list of n points spanning $\mathbb{R}^d$. Let μ be a distribution supported on $\mathcal{B}(U)$. For each $j \in [n]$, let $\gamma_j \in [0,1]$ be the probability that $j \in I$, when $I \subset [n]$ is sampled according to μ. Then $\gamma = (\gamma_1, \ldots, \gamma_n)$ is in $\mathcal{K}(U)$.*

Proof. The vector γ is easily seen to be equal to the convex combination

$$\sum_{I\in\mathcal{B}(U)} \mu(I) \cdot 1_I. \qquad \square$$

Theorem 5.4 ([6])**.** *Let $U = (u_1, \ldots, u_n) \in (\mathbb{R}^d)^n$ be a list of n points spanning $\mathbb{R}^d$ and let*

$$\gamma = (\gamma_1, \ldots, \gamma_n) \in \mathcal{K}(U)^o.$$

Then there exists a real invertible $d \times d$ matrix M such that, denoting $\hat{u}_j = Mu_j/\|Mu_j\|$, we have

$$\sum_{j=1}^{n} \gamma_j \cdot (\hat{u}_j \otimes \hat{u}_j) = I_{d\times d}. \tag{5.1}$$

[9] The relative interior of a set is a subset of the points of the set that are not on the boundary of the set, relative to the smallest subspace containing the set.

Proof. We will show how the proof follows from one of the propositions proved in [6] (whose proof we will not repeat here). The idea is to define a certain optimization problem parametrized by γ and to show that the maximum is achieved for all $\gamma \in \mathcal{K}(U)$. Then, the matrix M will arise from equating the gradient to zero at the maximum and solving the resulting equations.

We start by defining the optimization problem. For $t \in \mathbb{R}^n$ we define

$$X = X(t) = \sum_{j=1}^{n} e^{t_j} \cdot (u_j \otimes u_j) .$$

Notice that $X(t)$ has a positive determinant for all $t \in \mathbb{R}^n$, since U spans $\mathbb{R}^d$. Let $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined as

$$f(\gamma, t) = \langle \gamma, t \rangle - \ln \det(X(t)) .$$

The optimization problem is defined as

$$\phi^*(\gamma) = \sup_{t \in \mathbb{R}^n} f(\gamma, t) .$$

We now state a claim from [6] which gives sufficient conditions for the supremum $\phi^*(\gamma)$ to be realized.

Claim 5.5 (Rephrased from Proposition 6 in [6])**.** *If $\gamma \in \mathcal{K}(U)^o$ then the supremum $\phi^*(\gamma)$ is achieved. That is, there exists $t^* \in \mathbb{R}^n$ such that $f(\gamma, t^*) = \phi^*(\gamma)$.*

Let $t^* \in \mathbb{R}^n$ be a maximizer given by the claim. We can now use the fact that the partial derivatives

$$\frac{\partial f(\gamma, t)}{\partial t_j}$$

all vanish at the point t^*. Recall that

$$\frac{d}{ds} \ln \det(A) = \mathsf{tr}\left(A^{-1} \frac{d}{ds} A \right)$$

at all points where A is invertible [24, Ch. 9, Thm. 4]. Taking the derivative of f at t^* then gives

$$0 = \frac{\partial f(\gamma, t)}{\partial t_j}(t^*) = \gamma_j - \mathsf{tr}\left(X(t^*)^{-1} e^{t_j^*} (u_j \otimes u_j) \right) .$$

Since $X(t^*)^{-1}$ is positive definite, there exists a symmetric matrix M so that $M^2 = X(t^*)^{-1}$. Plugging this into the last equation and using properties of the trace function, we get

$$0 = \gamma_j - e^{t_j^*} \|M u_j\|^2 .$$

This means that

$$M^{-2} = X(t^*) = \sum_{j=1}^{n} \frac{\gamma_j}{\|M u_j\|^2} \cdot (u_j \otimes u_j) = \sum_{j=1}^{n} \gamma_j \cdot \left(\frac{u_j}{\|M u_j\|} \otimes \frac{u_j}{\|M u_j\|} \right) .$$

Multiplying by M from both sides we get

$$I_{d\times d} = \sum_{j=1}^{n} \gamma_j \cdot \left(\frac{Mu_j}{\|Mu_j\|} \otimes \frac{Mu_j}{\|Mu_j\|} \right)$$

as was required. □

Proof of Lemma 5.1. Let $\gamma \in \mathbb{R}^n$ be such that $\gamma_j = \Pr_{I\sim\mu}[j \in I]$ for all $j \in [n]$. By Claim 5.3, $\gamma \in \mathcal{K}(U)$. This means we can find $\gamma' \in \mathcal{K}(U)^o$ of distance at most ε from γ for all $\varepsilon > 0$. Hence, we can choose ε sufficiently small so that $\alpha/2 \leq \gamma'_j$ for all $j \in S$. Using Theorem 5.4 we get that there exists an invertible M so that

$$I_{d\times d} = \sum_{j=1}^{n} \gamma'_j \cdot (\hat{u}_j \otimes \hat{u}_j) .$$

Multiplying by the column vector w from the left and by the row vector w^t from the right we get that

$$1 = \langle w, w\rangle = \sum_{j=1}^{n} \gamma'_j \langle \hat{u}_j, w\rangle^2 \geq (\alpha/2) \sum_{j\in S} \langle \hat{u}_j, w\rangle^2 .$$

This completes the proof. □

6 Reducing to the well-spread vectors case

In this section we prove a lemma saying that, when analyzing an LCC $V = (v_1, \ldots, v_n)$ over $\mathbb{R}$, we can assume that the elements of V are unit vectors pointing in well-spread directions. The notion of well-spread vectors that we use is that given by Barthe's theorem (Lemma 5.1). More formally, the lemma will say that *any* list of vectors can be transformed into a list that is well spread as long as it does not contain a large subset of low rank. We formalize this result in Theorem 6.4. Below we state a lemma which basically follows as a corollary of the above theorem when the original list of vectors is an LCC. We first state and prove this lemma.

Lemma 6.1. *Let $V = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$ be a $(3,\delta)$-LCC be such that any subset $V' \subset V$ with $|V'| \geq (\delta/4)n$ satisfies $\dim(V') > 4\beta d$. Then, there exists a subset $U = (u_1, \ldots, u_{n'}) \subset V$ that is a $(3,\delta/2)$-LCC with $|U| = n' \geq (1-\delta/2)n$, and an invertible linear map $M : \mathbb{R}^d \to \mathbb{R}^d$ so that, denoting $\hat{u}_j = Mu_j/\|Mu_j\|$, we have for all unit vectors $w \in \mathbb{R}^d$.*

$$\sum_{j\in[n']} \langle \hat{u}_j, w\rangle^2 \leq \frac{n}{\beta d} .$$

Recall that (Observation 4.2) applying an invertible linear map to the elements of an LCC V preserves the property of being an LCC. Hence, if we are aiming to prove that a $(3,\delta)$-LCC V has a large subset of low rank we could use Lemma 6.1 to reduce to the case that the points of V are well spread.

We will prove Lemma 6.1 using Lemma 5.1. Recall that, Lemma 5.1 provides us with sufficient conditions under which a linear map M as in the lemma exists. Namely, that there exists a distribution μ on spanning d-tuples of V which hits each element in V with probability not too small. We will show that, if this condition does not hold (that is, if such a μ does not exist), we can find a large low-rank

subset V'. The high-level idea is to consider the greedy distribution on d-tuples that is sampled as follows: iteratively pick a random unspanned element from V and add it to the spanning set until we cover all of V. If this distribution gives low probabilities for many elements of V then we show that it must be due to the fact that these elements lie in some low-dimensional subspace. The following definition will be crucial for this argument.

Definition 6.2 ((η,τ)-independent set)**.** Let $U = (u_1,\ldots,u_n) \in (\mathbb{R}^d)^n$ be a list of n points spanning $\mathbb{R}^d$. We say that U is (η,τ)-independent, if there exists a distribution μ supported on $\mathcal{B}(U)$, and a set $S \subseteq [n]$ with $|S| \geq (1-\eta)n$ such that for all $j \in S$

$$\tau\frac{d}{n} \leq \Pr_{I\sim\mu}[u_j \in I]\,.$$

Since every $I \sim \mu$ has exactly d elements, observe that for every distribution μ,

$$E_j\Big[\Pr_{I\sim\mu}[u_j \in I]\Big] = d/n\,.$$

Moreover, if the points were in "general position," i. e., every d of the points were linearly independent, then by taking the distribution μ to be the uniform distribution on d-tuples with distinct elements, we would get a $(0,1)$-independent set.

Lemma 6.3. *Let $U = (u_1,\ldots,u_n) \in (\mathbb{R}^d)^n$. If U is not (η,τ)-independent, then there exists a subspace W of dimension at most τd which contains at least ηn elements of U.*

Proof. We construct a subset $S \subset [n]$ by the following greedy process. Start with $S = \emptyset$. At the jth step we check whether the vectors $\{u_i \mid i \notin S\}$ span a subspace of dimension at least τd. If they do, we add to S a tuple S_j of size $\lceil \tau d \rceil$ that is linearly independent. (That is, $\{u_i \mid i \in S_j\}$ are linearly independent vectors.) If $\{u_i \mid i \notin S\}$ have dimension lower than τd we halt. Let W be the subspace spanned by the complement of S at the end of this process. Notice that W has dimension at most τd.

Now, consider the following distribution on $\mathcal{B}(U)$. We first pick uniformly at random one of the sets S_j described above and add to our basis the corresponding (linearly independent) elements of U. Then we complete this set to a basis in some fixed way. For example, this can be done by taking the first basis in some fixed order that contains the elements of S_j. For each element in S, the probability of picking it to be in the basis is $\lceil \tau d \rceil / |S| \geq \tau d/n$. Since we are assuming that U is not (η,τ)-independent, the size of S^c must be at least ηn. By the definition of W, this completes the proof. □

Proof of Lemma 6.1. Applying Lemma 6.3 we get that V must be $(\delta/2, 2\beta)$-independent. Otherwise, V would contain a subset V' of size $(\delta/4)n$ and dimension at most $4\beta d$ (contradicting the assumption in the lemma). Hence, there exists a distribution μ on $\mathcal{B}(U)$ and a set $S \subset [n]$ with $|S| \geq (1-\delta/2)n$ such that for all $j \in S$

$$2\beta\frac{d}{n} \leq \Pr_{I\sim\mu}[j \in I]\,.$$

Let $U = V_S = \{v_i \mid i \in S\} = (u_1,\ldots,u_{n'})$ with $n' = |S|$. Lemma 5.1 now implies that there there exists an invertible linear map M so that, denoting $\hat{u}_j = Mu_j/\|Mu_j\|$, we have for all unit vectors $w \in \mathbb{R}^d$

$$\sum_{j\in S}\langle \hat{u}_j, w\rangle^2 \leq \frac{n}{\beta d}\,.$$

Notice that U is a $(3,\delta/2)$-LCC since the complement of U can intersect at most $\delta n/2$ triples from each matching in V. This completes the proof of the lemma. □

6.1 A convenient form of Barthe's theorem

The proof of Lemma 6.1 actually gives a more general result (not mentioning LCCs) that might be of independent interest.

Theorem 6.4. *Let $V=(v_1,\ldots,v_n)\in(\mathbb{R}^d)^n$ with $\dim(V)=d$ be so that any subset $U\subset V$ of size $|U|\geq\alpha n$ has $\dim(U)\geq\beta d$. Then, there exists an invertible linear map $M:\mathbb{R}^d\to\mathbb{R}^d$ and a subset $S\subset V$ of size $|S|\geq(1-2\alpha)n$ so that, if we denote by $\hat{v}=Mv/\|Mv\|$, we have for all unit vectors $w\in\mathbb{R}^d$*

$$\sum_{v\in S}\langle\hat{v},w\rangle^2\leq\frac{4n}{\beta d}.$$

Proof. The conditions on V and Lemma 6.3 imply that V is $(2\alpha,\beta/2)$-independent. Then, using Lemma 5.1, we get the map M and a set S as required. □

7 Reduction to the low triple-multiplicity case

In this section we prove a lemma that shows that, when analyzing a $(3,\delta)$-LCC V over any field $\mathbb{F}$, it is enough to consider codes in which the matchings $M_v, v\in V$ used in the decoding are such that each triple appears in a small number of matchings. (Otherwise we can find a large subset of low rank.)

Definition 7.1 (Triple-multiplicity)**.** We say that a $(3,\delta)$-LCC V with matchings $M_v, v\in V$ has *triple-multiplicity* at most r if each triple in each M_v appears in at most r of the matchings.

Lemma 7.2. *Let $\mathbb{F}$ be a field, $n\geq(1/\delta)^{\omega(1)}$ and $\beta>0$ a constant. Let $V=(v_1,\ldots,v_n)\in(\mathbb{F}^d)^n$ be a $(3,\delta)$-LCC with matchings $M_v, v\in V$. Suppose that any subset $V'\subset V$ with $|V'|>(\delta^2/36)n$ satisfies $\dim(V')>n^{1/2-\beta/4}$. Then, there exists a $(3,\delta/24)$-LCC $U\subset V$ with $|U|\geq(\delta/4)n$ and matchings $M'_v, v\in U$ so that U (with the matchings M'_v) has triple-multiplicity at most n^β and the matchings M'_v are subsets of the corresponding matchings M_v.*

Proof. We first reduce to the situation where every element participates in many triples. Unless mentioned otherwise, we will count triples with multiplicity. Let $0<\gamma=\delta^2/6$ be a real number. Iteratively delete vertices from V that participate in less than γn triples (counted with multiplicity), and the triples they participate in. Let $B\subseteq V$ be the subset of deleted elements, and let $V'=V\setminus B$. Since each deleted vertex only gets rid of γn triples, the total number of triples which include some vertex of B is at most γn^2. Thus each element in V' participates in at least γn triples, and at least $(\delta-\gamma)n^2>(2\delta/3)n^2$ of the triples in V are supported entirely in V'. Call this set of triples T'.

Claim 7.3. $|V'|>2\delta n$.

Proof. This is because there must be some $v\in V$ with at least $(2\delta/3)n$ triples in its matching that still survive in T'—if this was not the case, we would have $|T'|<(2\delta/3)n^2$. Since the triples in the matching corresponding to v are disjoint, $|V'|\geq 2\delta n$. □

Let $B' \subset V'$ be the subset of points in V' which have less than $\delta n/2$ of the triples in their matching supported in V'. Let $V'' = V' \setminus B'$.

Claim 7.4. *$|V''| \geq \delta n$, and V'' is a $(3,(\delta/6)(n/|V''|))$-LCC.*

Proof. There can be at most $\delta n/3$ elements in V' such that $\delta n/2$ triples in their matchings include an element from B—if there were more than that, then the total number of triples including a element from B would be greater than $\delta n/3 \cdot \delta n/2 \geq \delta^2 n^2/6 \geq \gamma n^2$, which is not possible. Thus, at least $|V'| - \delta n/3$ of the elements in V' have a matching of size at least $\delta n/2$ decoding them, lying wholly within V'. Thus $|B'| \leq \delta n/3$. Hence $|V''| \geq |V'| - |B'| \geq |V'| - \delta n/3 > \delta n$. Moreover, for each $v \in V''$, it has a matching of size at least $\delta n/2 - |B'| \geq \delta n/6$ supported in V''. Thus V'' is a $(3,(\delta/6)(n/|V''|))$-LCC. Let T'' be the union of all the triples in the LCC V''. □

We will call a triple in T'' a *high-multiplicity* triple if it has multiplicity at least n^β in T''; otherwise we will call it a *low-multiplicity triple*).

Claim 7.5. *At least $(1-\delta/24)|V''|$ of the elements in V'' have a matching of size $(\delta/12)|V''|$ of low-multiplicity triples decoding them.*

Proof. Suppose the claim does not hold. That is, at least $(\delta/24)|V''|$ of the elements in V'' have at least half of their matchings (in T'') composed of high multiplicity triples.

We now delete all the triples of low multiplicity from T''. Since there are at least $(\delta^2/288)|V''|^2$ triples (counting multiplicity) of multiplicity at least n^β in the LCC V'', by averaging, there exists $v \in V''$ that participates in at least $(\delta^2/288)|V''|$ triples (counted with multiplicity), and each of the triples has multiplicity at least n^β. Observe that since all these triples contain v, no two triples are part of a matching corresponding to the same element.

By greedily choosing distinct triples containing v of highest multiplicity, one can pick a set T^* of distinct triples of size at most $n^{1/2-\beta/2}$ such that together they span at least $n^{1/2+\beta/2}$ distinct elements of V''. This is true since $n^{1/2+\beta/2} \leq (\delta^2/288)|V''|$, each triple of multiplicity n^β spans at least n^β distinct elements and distinct triples sharing an element must span distinct elements. By "distinct" here we mean distinct LCC indices (not necessarily distinct vectors).

Let L be a linear transformation of co-rank at most $3n^{1/2-\beta/2}$ which maps each element participating in a triple of T^* to 0. Since all the elements spanned by the triples of T^* also get mapped to 0, at least $n^{1/2+\beta/2}$ elements of V'' get mapped to 0 under L. Let this set be V^*. Recall that each element of V' (and hence of V^*) participates in γn triples which together decode γn distinct elements of V.

Let $S \subset V$ be the subset of all elements whose matching contains at least $(\gamma/6)n^{1/2+\beta/2}$ triples that each contain some element from V^*. Since the total number of triples containing some element from V^* is at least $|V^*| \cdot \gamma n/3$, by a simple counting argument we get that $|S| \geq (\gamma/6)n$.

To finish the proof of Claim 7.5 we will now argue that

$$\dim(S) \leq 2n^{1/2-\beta/3} < n^{1/2-\beta/4}.$$

For contradiction, assume $\dim(S) > 2n^{1/2-\beta/3}$, then

$$\dim(L(S)) > 2n^{1/2-\beta/3} - 3n^{1/2-\beta/2} > n^{1/2-\beta/3}.$$

Moreover, since L sends V^* to 0, all triples containing some element of V^* now have at most 2 nonzero elements, and thus the triples can be replaced by *pairs*. Thus $L(V)$ is a $(2,(\gamma/6)n^{-1/2+\beta/2})$-LDC of size n, decoding to linearly independent vectors spanning at least $n^{1/2-\beta/3}$ dimensions. Using Theorem 4.5 (lower bound for 2-query LDCs) we get that

$$n \geq 2^{\frac{\gamma/6n^{\beta/6}}{16}-1}.$$

Since $n \geq (1/\delta)^{\omega(1)}$, $\gamma = \mathsf{poly}(\delta)$ and $\beta = \Omega(1)$, this is a contradiction (for large enough n).

Thus, the set S has size at least $(\gamma/6)n = \delta^2 n/36$ and dimension at most $n^{1/2-\beta/4}$, contradicting the assumption in Lemma 7.2. This completes the proof of Claim 7.5 □

Applying Claim 7.5, we see that one can delete all triples of multiplicity greater than n^β and delete at most $\delta|V''|/24$ elements to get a subset U such that each element of U has a matching of $\delta|U|/24$ triples decoding to it where the triples are supported in U. Thus U is a $(3,\delta/24)$-LCC with $|U| \geq \delta n/4$, and with all triples of multiplicity at most n^β. This completes the proof of Lemma 7.2. □

8 LCCs over $\mathbb{R}$ can be clustered

In this section we prove the "clustering step" described in the introduction.

Definition 8.1 (Clustering)**.** Let $S_1,\ldots,S_m \subset [n]$. We say that a triple $\tau \in \binom{[n]}{3}$ is *clustered* by the family of sets $S_1,\ldots,S_m$ if there exists an $i \in [m]$ so that $|\tau \cap S_i| \geq 2$. If M is a multiset of triples, we say that M is clustered by $S_1,\ldots,S_m$ if every triple in M is clustered.

We prove the clustering result as a sequence of three lemmas. First we state the final clustering lemma that will be used later in the proof of our main result. This result was informally stated as Lemma 3.1.

Lemma 8.2 (Final clustering)**.** *Let $n > (1/\delta)^{\omega(1)}$ and let $\beta > 0$ be a constant. Let $V = (v_1,\ldots,v_n) \in (\mathbb{R}^d)^n$ be a $(3,\delta)$-LCC so that every subset $U \subset V$ of size $|U| \geq (\delta^2/288)n$ has dimension at least $\max\{8\delta^6 d, n^{1/2-\beta/4}\}$. Then, there exists a $(3,\hat{\delta})$-LCC $\hat{V} = (\hat{v}_1,\ldots,\hat{v}_{\hat{n}}) \subset V$ of dimension $\hat{d} \leq d$, size $\hat{n} \geq (\delta/10)n$ and $\hat{\delta} \geq \delta^2/4$ and sets $S_1,\ldots,S_m \subset [\hat{n}]$ so that*

1. *$|S_i| \leq O(\hat{n}/\hat{\delta}^6\hat{d})$ for all $i \in [m]$;*
2. *$\Omega(\hat{\delta}^{19}\hat{d}^3/\hat{n}^{1+2\beta}) \leq m \leq O(\hat{n}^{1+2\beta}/\hat{\delta}^{10}\hat{d})$;*
3. *if $\hat{M}_{\hat{v}}, \hat{v} \in \hat{V}$ are the matchings used to decode $\hat{V}$, then each $\hat{M}_{\hat{v}}$ is clustered by $S_1,\ldots,S_m$.*

We will prove this lemma using the following lemma, which adds conditions on the given code.

Lemma 8.3 (Intermediate Clustering)**.** *Let $n \geq (1/\delta)^{\omega(1)}$ and $\beta > 0$ a constant. Let $V = (v_1,\ldots,v_n) \in (\mathbb{R}^d)^n$ be a $(3,\delta)$-LCC with triple-multiplicity at most n^β and so that for each unit vector $w \in \mathbb{R}^d$*

$$\sum_{j=1}^{n} \langle v_j, w\rangle^2 \leq \frac{n}{\delta^6 d}.$$

Let $t = n/(\delta^6 d)$ and suppose that $d > 10^8 \cdot 200/\delta^8$. Then, there exist m subsets $S_1,\ldots,S_m \subset V$ such that

1. $|S_i| \leq O(t)$ *for all* $i \in [m]$;
2. $\Omega(\delta n^{2-\beta}/t^3) \leq m \leq O(t \cdot n^{\beta}/\delta^4)$;
3. *if* $M = \bigcup_{v \in V} M_v$ *is the multiset of all triples in all matchings used to decode* V*, then there are at most* $\delta^2 n^2/100$ *triples in* M *that are not clustered by* $S_1, \ldots, S_m$.

To prove the intermediate clustering lemma we first prove a basic clustering lemma.

Lemma 8.4 (Basic Clustering)**.** *Let* n, t, β, δ *and* $V \in (\mathbb{R}^d)^n$ *be as in Lemma 8.3 and let* M *be the multiset of triples obtained by taking the union of all* $M_v, v \in V$*. Let* $\bar{M} \subset M$ *be of size at least* $\delta^2 n^2/100$ *and suppose that* $d > 10^8 \cdot 200/\delta^8$*. Then there exists a subset* $S \subset V$ *with* $|S| \leq O(t)$ *and a subset* $T \subset \bar{M}$ *with* $|T| \geq \Omega(\delta^4 n^{2-\beta}/t)$ *such that each triple in* T *contains at least two elements from* S.

First, we show how to use the intermediate clustering lemma to prove the final Lemma 8.2. After that, we will prove the basic clustering lemma and, using it, easily derive the intermediate clustering lemma.

Proof of Lemma 8.2. At a high level, the proof follows by first applying Lemma 6.1 to get the well-spread vectors condition on the points in a large sub-LCC V' of V. Then we use Lemma 7.2 on V' to get a subcode V'' with low triple-multiplicity. (This does not ruin the well-spread vectors condition by much.) Finally, we apply Lemma 8.3 on V'' to get clustering for almost all triples. The only reason why one of these steps could fail is if we found a large low dimensional subset in V (which will contradict our assumptions). A final refinement step, using Claim 4.1 shows the existence of a subcode $\hat{V}$ as required. The details follow.

Reducing to the well-spread vectors case. We apply Lemma 6.1 on V, with $\beta = 2\delta^6$, to obtain a subset V' of size $n' \geq (1-\delta/2)n$ so that V' is a $(3, \delta' = \delta/2)$-LCC and so that for each unit vector $w \in \mathbb{R}^d$ we have

$$\sum_{v' \in V'} \langle v', w \rangle^2 \leq \frac{n}{2\delta^6 d}.$$

If we cannot apply Lemma 6.1, it means that there is a subset U in V of size $|U| \geq (\delta/4)n$ and dimension at most $8\delta^6 d$, which would contradict our assumptions.

Reducing to low triple-multiplicity. We now apply Lemma 7.2 on the LCC V' to get a $(3, \delta/48)$-LCC $V'' \subset V'$ of size $n'' \geq (\delta/8)n$ and with triple-multiplicity at most $(n')^{\beta} \leq (n'')^{2\beta}$. If we cannot apply the lemma, it means that there is a subset $U \subset V'$ of size $|U| \geq (\delta'^2/36)n' \geq (\delta^2/288)n$ and dimension $\dim(U) \leq (n')^{1/2-\beta/4} \leq n^{1/2-\beta/4}$, which would contradict our assumptions. Let $d'' = \dim(V'')$ and $\delta'' = \delta^2/2$. We can think of V'' as a $(3, \delta'')$-LCC over $\mathbb{R}^{d''}$ in which the well-spread vectors condition above can be written as

$$\sum_{v'' \in V''} \langle v'', w \rangle^2 \leq \frac{n}{2\delta^6 d} \leq \frac{n''}{\delta''^6 d''},$$

for all unit vectors $w \in \mathbb{R}^{d''}$. (We took $\delta'' = \delta^2/2$ to compensate for the drop in n'' in the above inequality.) Notice that moving from $\mathbb{R}^d$ to $\mathbb{R}^{d''}$ is not a problem since we can orthogonally project all vectors on the span of V'' and maintain all inner products with all unit vectors.

Clustering. We can now apply Lemma 8.3 on V'' to find sets $S_1,\ldots,S_m$ that cluster all but $(\delta''^2/100)n''^2$ of the triples in the decoding matchings of V''. With $|S_i| \le O(n''/\delta''^6 d'')$ for all $i \in [m]$ and (using $t = n''/\delta''^6 d''$) we obtain

$$\Omega\left(\frac{\delta''^{19} d''^3}{n''^{1+2\beta}}\right) \le m \le O\left(\frac{n''^{1+2\beta}}{\delta''^{10}} \cdot d''\right).$$

If we cannot apply the lemma, it means that $d'' \le (1/\delta'')^{O(1)}$, which would contradict our assumptions on V (since it would have a subset V'' of size $n'' \ge (\delta/8)n$ and dimension $(1/\delta)^{O(1)} < n^{1/4}$).

Refinement. To complete the proof, observe that, there are at least $(1-\delta''/10)n''$ points in V'' that have at least half of their matchings clustered by $S_1,\ldots,S_m$. Hence, we can use Claim 4.1 to find a $(3,\hat{\delta})$-LCC $\hat{V} \subset V''$ of size $\hat{n} \ge (1-\delta''/10)n'' \ge (\delta/10)n$ with $\hat{\delta} \ge \delta''/2 \ge \delta^2/4$ so that the sets $S_1,\ldots,S_m$ (restricted to indices in $\hat{V}$) cluster all the triples in the matchings of $\hat{V}$. Notice that, since $\hat{d} = d''$, $\hat{\delta} = \Theta(\delta'')$ and $\hat{n} = \Theta(n'')$, the bounds on the sizes of the sets S_i and on m still hold (the difference in constants will be swallowed by the big "O"). This completes the proof of Lemma 8.2. □

8.1 Preliminaries for the proof of the clustering lemmas

We denote by $\|v\|$ the ℓ_2 norm of a vector v. Notice that for two unit vectors u and v, $\|u-v\|^2 = 2-2\langle u,v\rangle$. We denote the *correlation* between two unit vectors v,u as $|\langle v,u\rangle|$.

Let V be as in Lemma 8.3 with matchings $M_v, v \in V$. The conditions of Lemma 8.3 (which we will assume to hold for the rest of this section) tell us that for all unit vectors $u \in \mathbb{R}^d$ we have

$$\sum_{j=1}^{n} \langle v_j,u\rangle^2 \le \frac{n}{\delta^6 d} = t\,. \tag{8.1}$$

This has the following useful consequence.

Claim 8.5. *For every unit vector $u \in \mathbb{R}^d$ we have*

$$|\{v \in V \mid |\langle v,u\rangle| \ge \alpha\}| \le t/\alpha^2\,.$$

We can also bound the number of points in V that correlate with a given plane.

Claim 8.6. *Let $P \subset \mathbb{R}^d$ be a two dimensional subspace and define*

$$K = \{v \in V \mid |\langle v,u\rangle| \ge \alpha \text{ for some unit vector } u \in P\}\,.$$

Then $|K| \le (80/\alpha^3)\cdot t$.

Proof. For each $v \in K$ let $u(v) \in P$ be a unit vector with $|\langle v,u(v)\rangle| \ge \alpha$. Now, cover the boundary of the unit circle in P with at most $20/\alpha$ balls[10] of radius at most $\alpha/2$. By a pigeon hole argument, one of these balls must contain at least $\alpha|K|/20$ of the points $u(v)$. Now, the center of this ball must have correlation at least $\alpha/2$ with all the $\alpha|K|/20$ corresponding vectors v. Applying Claim 8.5 we get that $|K| \le (80/\alpha^3)t$. □

[10] We consider balls in $\mathbb{R}^d$.

For every unit vector $u \in \mathbb{R}^d$, let

$$\mathsf{Cor}(u) = \{v \in V \mid |\langle u, v\rangle| \geq 1/10^4\}.$$

For every $v \in V$, let $M_v^* \subseteq M_v$ be defined as

$$M_v^* = \{(v_i, v_j, v_k) \in M_v \mid v_i, v_j, v_k \in V \setminus \mathsf{Cor}(v)\}.$$

That is, let M_v^* be the subset of the triples decoding v where each vector in each triple has low correlation with v. Intuitively, such triples must be close to a two dimensional plane and hence "almost" dependent.

The following is an immediate corollary of Claim 8.5.

Claim 8.7. *For every $v \in V$, $|M_v^*| \geq |M_v| - 10^8 t \geq \delta n - 10^8 t$.*

Let M^* be the (multiset) union of all triples in M_v^* for all $v \in V$. By Claim 8.7, M^* has size at least $\delta n^2 - 10^8 tn$.

The following proposition bounds the number of triples in M^* containing a fixed pair of vertices.

Proposition 8.8. *For all $i \neq j \in [n]$, there are at most $O(tn^\beta)$ triples (counting multiplicities) in M^* containing the pair (v_i, v_j).*

Proof. We will show a bound of $O(t)$ on the number of *distinct* triples containing (v_i, v_j). The $O(tn^\beta)$ bound will then follow by our assumption on the maximum multiplicity of triples in M (and so also in M^*).

Let $P = \mathsf{span}\{v_i, v_j\}$. Consider a triple (v_i, v_j, v_k) containing v_i, v_j and suppose this triple belongs to some matching M_v^*. Let $\Pi = \mathsf{span}\{v_k, v\}$ and observe that both planes P and Π (both are indeed planes since the property of the LCC being regular implies the distinctness of the points in a triple and the point they are used to decode to) are contained in the three dimensional subspace $\mathsf{span}\{v_i, v_j, v_k\}$. Therefore, they must intersect in some unit vector $w \in P \cap \Pi$. Now, since $|\langle v_k, v\rangle| \leq 10^{-4}$, a simple calculation shows that w must have correlation at least $1/10$ with either v_k or v (since w belongs to their span and they are close to being orthogonal). To summarize, we have shown that in every triple $(v_i, v_j, v_k) \in M_v$, one of the vectors v, v_k has correlation at least $1/10$ with the plane P. Now, the union of $\{v, v_k\}$ as we go over all distinct triples containing $\{v_i, v_j\}$ is at most $O(t)$ by Claim 8.6. If the total number of distinct triples is r, then at least $r/2$ of the vectors v will correlate with P or $r/2$ of the v_k will correlate with P. In either case we see that $r/2 = O(t)$, and hence $r = O(t)$. □

Definition 8.9 (Triple types)**.** We split the triples appearing in M^* into two *types*.

- A triple $(v_i, v_j, v_k) \in M^*$ is defined to be of *Type A* if there exists a pair of vertices in the triple, say (v_i, v_j), such that
$$|\langle v_i, v_j\rangle| \geq 9/10.$$
- A triple $(v_i, v_j, v_k) \in M^*$ is defined to be of *Type B* if
$$|\langle v_i, v_j\rangle| < 9/10, \qquad |\langle v_j, v_k\rangle| < 9/10 \quad \text{and} \quad |\langle v_i, v_k\rangle| < 9/10.$$

When we refer to a triple as Type A or Type B, we will implicitly assume that this triple is in M^*.

We first state and prove three simple propositions that will be useful in the proof of the basic clustering lemma. Below, we will sometimes refer to the elements of V as "vertices." The reader not wishing to follow the somewhat tedious calculations can recall the high level overview given in Section 3.

Proposition 8.10. *Let (v_i, v_j, v_k) be a triple of Type B then either $|\langle v_i, v_j\rangle| \geq 1/100$ or $|\langle v_i, v_k\rangle| \geq 1/100$.*

Proof. Suppose in contradiction that $\langle v_i, v_j\rangle < 1/100$ and $\langle v_i, v_k\rangle < 1/100$.

Suppose the triple decodes to the vector u and by an appropriate orthogonal change of basis (which does not change distances or inner products), let us assume that the vectors all lie in the 3 dimensional space spanned by the unit vectors e_1, e_2 and e_3. We can also assume that $u = e_1$, v_i is a linear combination of e_1 and e_2, and v_j and v_k are linear combinations of e_1, e_2 and e_3.

Since the vectors in the triple are uncorrelated to u, their inner product with e_1 has absolute value at most $1/10^4$. Since v_i is a unit vector, $\langle v_i, e_1\rangle^2 + \langle v_i, e_2\rangle^2 = 1$ and hence $|\langle v_i, e_2\rangle| > |\langle v_i, e_2\rangle|^2 \geq 1 - 1/10^8$.

Also since $|\langle v_i, v_j\rangle| < 1/100$ and $|\langle v_i, v_k\rangle| < 1/100$,

$$|\langle v_j, e_2\rangle| < \frac{1}{100} \times \frac{10^8}{10^8 - 1} < \frac{2}{100}.$$

Similarly $|\langle v_k, e_2\rangle| < 2/100$. Also since v_j is a unit vector, $\langle v_j, e_1\rangle^2 + \langle v_j, e_2\rangle^2 + \langle v_j, e_3\rangle^2 = 1$ and hence $\langle v_j, e_3\rangle^2 \geq 1 - 1/10^8 - 4/10^4$, implying that $|\langle v_j, e_3\rangle| \geq \sqrt{99/100}$. Similarly $|\langle v_k, e_3\rangle| \geq \sqrt{99/100}$. Hence $|\langle v_k, v_j\rangle| \geq 99/100$, contradicting the property of being Type B. □

Proposition 8.11. *Suppose T is a set of m* distinct *triples of Type B, each sharing the pair (v_i, v_j). Let S be the set of size m containing all the vertices of the triples in T except v_i and v_j. Then there is a ball of radius at most $5/10^4$ containing at least $m/10^5$ points of S.*

Proof. We will first show that every point of S is close to the subspace through v_i and v_j, and then apply a pigeon hole argument.

Let $v_k \in S$. Then (v_i, v_j, v_k) is a triple of Type B, and in particular the triple is in M_u^* for some vertex u.

By an appropriate orthogonal change of basis (which does not change distances or inner products), we can assume that the vectors all lie in the 3 dimensional space spanned by the unit vectors e_1, e_2 and e_3. We can also assume that $v_i = e_1$, v_j is a linear combination of e_1 and e_2, and u and v_k are linear combinations of e_1, e_2 and e_3.

Since we have a triple of Type B, $|\langle v_i, v_j\rangle| < 9/10$. Thus $|\langle v_j, e_1\rangle| < 9/10$. Since $\langle v_j, e_1\rangle^2 + \langle v_j, e_2\rangle^2 = 1$, this implies that $|\langle v_j, e_2\rangle| > 2/5$. Also since $|\langle u, v_i\rangle| < 1/10^4$ and $|\langle u, v_j\rangle| < 1/10^4$, thus $|\langle u, e_1\rangle| < 1/10^4$ and $|\langle u, e_2\rangle| < 5/2 \times |\langle u, v_j\rangle| < 5/2 \times 1/10^4$. Hence

$$|\langle u, e_3\rangle| = \sqrt{1 - |\langle u, e_1\rangle|^2 - |\langle u, e_2\rangle|^2} \geq 1 - 1/10^7.$$

Since $|\langle u, v_k\rangle| < 1/10^4$, we get that $|\langle v_k, e_3\rangle| \leq 1/10^4 \times 10^7/(10^7 - 1) \leq 2/10^4$. Notice that $|\langle v_k, e_3\rangle|$ is precisely the distance of v_k to the subspace spanned by v_i and v_j.

Now consider the unit circle C in the subspace spanned by e_1 and e_2. We will show that each element of S is at distance at most $4/10^4$ from C. To see this, observe that for $v_k \in S$, the projection $\bar{v}_k$ of v_k onto

the subspace spanned by e_1 and e_2 is of length at least $1-2/10^4$ (by the triangle inequality). Thus $\bar{v}_k$ is at distance at most $2/10^4$ from C and also at distance at most $2/10^4$ from v_k. Thus again by the triangle inequality, the distance between v_k and C is at most $4/10^4$. Now cover C with 10^5 2-dimensional discs of radius $1/10^4$. Clearly this can be done. Thus each element v_k in S is at distance at most $5/10^4$ from the center of one of these discs. Thus for one of these discs, there are $m/10^5$ points of S that are at distance at most $5/10^4$ from the center of the disc. □

Proposition 8.12. *Let G be an edge-weighted k-uniform hypergraph on n vertices with $k \geq 2$. Define the degree of a vertex to be the sum of the weights of all hyperedges containing it. Suppose the average degree of a vertex in G is D. Then, there exists a vertex induced subgraph G' of G in which every vertex has degree at least D/k.*

Proof. To obtain G' we iteratively delete vertices whose degree in G is less than D/k. Observe that, after each deletion, the average degree in the hypergraph strictly increases. Indeed, after removing a vertex of degree $D' < D/k$, the new average is

$$\frac{n \cdot D - kD'}{n-1} > D\,.$$

Thus the process must terminate when all vertices have degree at least D/k. □

8.2 Basic clustering: proof of Lemma 8.4

At this point, we assume that V is well spread over the unit sphere, has low-multiplicity triples that are nearly orthogonal to their associated vectors in V. Each triple is either of Type A—containing a very correlated pair—or of Type B—consisting of uncorrelated vectors.

We first show that having many triples of the same type implies that we can find a small set of vertices such that many of the triples intersect the set in at least two of their elements. This will be the main step in the proof of Lemma 8.4 which is given below. Recall that we have an upper bound of n^β on the multiplicity of each triple in M^*.

Lemma 8.13. *Suppose there is a subset T of γn^2 triples (counting multiplicities) in M^* of the same type (either Type A or Type B), then there is a set $S \subseteq V$ such that $|S| = O(t)$, and at least $\Omega(\gamma^2 n^{2-\beta}/t)$ triples in T intersect S in at least two of their elements.*

Proof. We separate into two cases according to the type of the triples in T. In both cases, we will first refine to the situation where every vertex is incident to many (γn) triples. In both cases we will find a cluster V^* of nearby vertices, and let S be some kind of neighborhood of V^* such that every triple which intersects V^* will also intersect S in two elements. Since V^* will be incident to many triples, we will conclude that many triples intersect S in two elements. Moreover we will ensure that every vertex in S will have some constant *correlation* with some fixed carefully chosen vertex w. Since every element in S correlates with vertex w, Claim 8.5 implies that S cannot be too large. In the case of Type A triples, the argument is fairly straightforward, whereas in the case of Type B triples the argument is more delicate.

Case 1: T has only triples of Type A. Consider the following weighted graph H on vertex set V in which the edges are all pairs v_i, v_j with $|\langle v_i, v_j\rangle| \geq 9/10$ and the weight of an edge (v_i, v_j) is the number of triples in T, counting multiplicities, that contain this pair (we can discard edges of weight zero). We define the degree of a vertex $\deg(v)$ as the sum of weights over all edges of H that contain v. Since $(1/2)\sum_v \deg(v) \geq |T|$ we have that the average degree in H is at least $D = 2|T|/n \geq 2\gamma n$.

Let H' be a vertex induced subgraph of H in which every vertex has degree at least $D/2$ (such a subgraph exists by Proposition 8.12). Let w be any vertex in H' and observe that, by Proposition 8.8, w must have at least $r = \Omega(\gamma n^{1-\beta}/t)$ distinct neighbors $u_1, \ldots, u_r$ (since the maximal weight of an edge is $O(tn^{\beta})$). Let $V^* = \{u_1, \ldots, u_r\}$. We define the set S to contain these vertices $u_1, \ldots, u_r \in V^*$ as well as all of their neighbors.

First, we argue that S cannot be too large. To see this, observe that, if (v_i, v_j) is an edge in H then v_j must have ℓ_2 distance at most $1/\sqrt{5}$ from either v_i or $-v_i$. Thus, since all vertices in S are at (graph) distance less than two from w, we have that they are all contained in the union of two balls of radius $2/\sqrt{5}$ around w and around $-w$. This means that all points in S must have correlation at least $4/6$ with w. Using Claim 8.5 we get that $|S| \leq O(t)$.

To see that there are many triples with two elements in S observe that the sum over all weights of edges touching $u_1, \ldots, u_r$ is at least $r \cdot \gamma n \geq \Omega(\gamma^2 n^{2-\beta}/t)$ (using the fact that H' has high minimum degree). Since every triple is counted at most 3 times in this sum we conclude that there are at least $\Omega\left(\gamma^2 n^{2-\beta}/t\right)$ triples with a pair in S.

Case 2: T has only triples of Type B. Consider the following 3-regular weighted hypergraph G. The set of vertices of G is the set V. For each triple $(v_i, v_j, v_k) \in T$ we have a hyper-edge in G with weight equal to the multiplicity of that triple in T. By Proposition 8.12, there is a subgraph G' of G such that every vertex of G' is incident to at least $\gamma n/3$ triples (counting weights) lying within G'.

Pick any vertex $v \in G'$. Let C_v be the multiset $\{v' \in V \mid |\langle v, v'\rangle| > 1/100\}$ of vectors that are slightly correlated with v. By Claim 8.5 (well-spread vectors condition) we have $|C_v| < t \cdot 10^4$. By Proposition 8.10, every triple containing v has another vertex v' such that $|\langle v, v'\rangle| > 1/100$ (and thus $v' \in C_v$). Thus by a simple averaging argument, it must be that for some $v' \in C_v$, the pair (v, v') participates in at least $\gamma n/(3|C_v|)$ triples (counting multiplicities). Using the bound on triple-multiplicity, we get that there is a *set* T^* of at least $\gamma n/(3|C_v|n^{\beta})$ distinct triples containing v and v'. Thus

$$|T^*| \geq \frac{\gamma n}{3|C_v|n^{\beta}} \geq \frac{\gamma n^{1-\beta}}{10^4 \cdot t}$$

and, by Proposition 8.11, at least $|T^*|/10^5$ vertices (of G') lie in a ball of radius $5/10^4$. Call this set of vertices V^*. Thus what we have so far is a set V^* of vertices of G' all lying in a ball of radius $5/10^4$, where

$$|V^*| \geq \frac{\gamma n^{1-\beta}}{3 \cdot 10^9 \cdot t}.$$

Recall that every point v_k in V^* is incident to at least $\gamma n/3$ triples lying within G', and, by Proposition 8.10, for each of the triples there exists a vertex v'_k distinct from v_k in that triple such that $|\langle v_k, v'_k\rangle| > 1/100$.

Let $S = \{u \in V \mid \exists w \in V^* \text{ s.t. } |\langle u, w\rangle| > 1/100\}$ be the set of all vertices that have correlation at least $1/100$ with some vertex of V^*. Fix $w \in V^*$. Then for any $u \in S$, by definition of S, there exists $w' \in V^*$ such that $\langle u, w'\rangle > 1/100$. Also, since radius of V^* is at most $5/10^4$ we have $\|w - w'\| \leq 1/10^3$. Together, these imply that $|\langle u, w\rangle| > 1/10^3$. Since this holds for all $u \in S$ (and for the same fixed w), by Claim 8.5 we get that $|S| < 10^6 t$.

Moreover, observe that each triple that intersects V^* must intersect S in two elements. Since each tripe in V^* is incident to at least $\gamma n/3$ triples, and each triple is counted at most 3 times, there must be at least

$$\Omega\left(\gamma n \times \frac{\gamma n^{1-\beta}}{10^9 \cdot t}\right) = \Omega(\gamma^2 n^{2-\beta}/t)$$

triples with a pair in S. □

Proof of Lemma 8.4. Since $d > 200 \cdot 10^8/\delta^8$ we have that

$$t = \frac{n}{\delta^6 d} < \frac{\delta^2 n}{200 \cdot 10^8}.$$

Thus, by Claim 8.7 we have that for each $v \in V$

$$|M_v \setminus M_v^*| \leq 10^8 t \leq \delta^2 n/200.$$

So, the set $\bar{M}^* = \bar{M} \cap M^*$ must have size at least $|\bar{M}| - \delta^2 n^2/200 \geq \delta^2 n^2/200$ triples. At least half of these triples are of the same type (A or B) and so we can apply Lemma 8.13 with $\gamma = \delta^2/400$ to get the required sets S and triples T. □

8.3 Intermediate clustering: proof of Lemma 8.3

We prove Lemma 8.3 by iteratively applying Lemma 8.4 until we have gathered "enough" clustered triples, where we call a triple "clustered" if it has intersection size at least 2 with one of the sets S_i.

We start with $\bar{M} = M$, which is initially of size $|\bar{M}| \geq \delta n^2 \geq \delta^2 n^2/100$. Applying Lemma 8.4 we get sets $T_1 \subset \bar{M}$ and $S_1 \subset V$ with $|S_1| \leq O(t)$ and so that all triples in T_1 are clustered. We now let $\bar{M} = \bar{M} \setminus T_1$ and continue in this manner to generate $S_2, S_3, \ldots, S_m$ and (disjoint) $T_2, T_3, \ldots, T_m$, removing the triples in the T_i from $\bar{M}$ as we proceed, until there are at most $\delta^2 n^2/100$ triples in M that are not clustered.

This only leaves the task of bounding the number of iterations, m. The upper bound follows from the fact that the sets T_i are disjoint, each of size at least $\Omega(\delta^4 n^{2-\beta}/t)$ and that $|M| \leq \delta n^2$. The lower bound follows from the observation that, by Proposition 8.8, each T_i can have size at most $|S_i|^2 \cdot O(t \cdot n^\beta) = O(n^\beta t^3)$. Since the union of the T_i contains at least $\Omega(|M|) \geq \Omega(\delta n^2)$ triples, we get that $m \geq \Omega(\delta n^{2-\beta}/t^3)$. This completes the proof of Lemma 8.3.

9 Clustering implies large low-rank subset

The main result of this section is the following lemma that gives a dimension upper bound for LCCs in which the triples are "clustered." An informal statement of this lemma was given as Lemma 9.1.

Notice that this lemma works over any field $\mathbb{F}$.

Lemma 9.1 (Clustering implies large low-rank subset)**.** *Let $\mathbb{F}$ be a field, $0 < \varepsilon < 1/50$, $0 < \beta < \varepsilon/2$ and suppose $n > (1/\delta)^{\omega(1)}$. Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}^d)^n$ be a $(3,\delta)$-LCC with matchings $M_v, v \in V$. Suppose there exist sets $S_1, \ldots, S_m \subset [n]$ with*

1. *$|S_i| \leq O(n/\delta^6 d)$ for all $i \in [m]$;*
2. *$\Omega(\delta^{19} d^3/n^{1+\beta}) \leq m \leq O(n^{1+\beta}/\delta^{10} d)$;*
3. *every triple in each M_v is clustered by $S_1, \ldots, S_m$.*

Then there is a subset $V' \subset V$ of size $|V'| \geq (\delta/2)n$ and rank at most $n^{1/2-\varepsilon}$.

This lemma will be an easy corollary to the following lemma, which shows that there is a small subset in V so that, when projecting this set to zero, the dimension of V drops by a lot.

Lemma 9.2 (Restriction lemma)**.** *Let n, β, ε, V and $S_1, \ldots, S_m$ satisfy the conditions of Lemma 9.1. Assume further that the matchings M_v are in regular form (no "2-query" triples). If $d > n^{1/2-\varepsilon}$ then there exists a subset $U \subset V$ with*

$$|U| \leq n^{1/4+7\varepsilon}$$

such that, if $\mathcal{L} : \mathbb{F}^d \to \mathbb{F}^d$ is any linear map with $U \subset \ker(\mathcal{L})$ then $\mathcal{L}(V) = \{\mathcal{L}(v) \mid v \in V\}$ is contained in a subspace of dimension at most $n^{10\varepsilon}$

We prove the Restriction lemma (Lemma 9.2) below, following the short proof of Lemma 9.1 from Lemma 9.2.

Proof of Lemma 9.1. Using Claim 4.7 we can reduce to the case that the code V and the matchings M_v are in regular form (that is, there are no "2-query" triples). Indeed, replacing V with the code given in Claim 4.7 leaves us with a new code (with n and δ the same up to a constant) satisfying the same clustering requirements (using the same sets $S_1, \ldots, S_m$) and with the same dimension. If we cannot apply Claim 4.7 it is because there is a subset $U \subset V$ of size $(\delta/2)n$ and dimension at most $O((1/\delta)\log(n)) < n^{1/2-\varepsilon}$, in which case the proof is done.

Next, suppose in contradiction that $d > n^{1/2-\varepsilon}$ (otherwise we let $V' = V$). Apply Lemma 9.2 to get a subset $U \subset V$ with $|U| \leq n^{1/4+7\varepsilon}$, such that, if we send U to zero by a linear map, the dimension of $\mathsf{span}\{V\}$ goes down to at most $n^{10\varepsilon}$. The existence of such a U implies that

$$d = \dim(V) \leq |U| + n^{10\varepsilon} \leq n^{1/4+7\varepsilon} + n^{10\varepsilon}$$

which gives a contradiction if $\varepsilon < 1/50$. □

9.1 Proof of Lemma 9.2

Using the assumption $d > n^{1/2-\varepsilon}$ we get that for each $i \in [m]$,

$$|S_i| = O(\delta^{-6} n^{1/2+\varepsilon})$$

and the number of sets, m, is between

$$\Omega(\delta^{19} n^{1/2-3\varepsilon-\beta}) \leq m \leq O(\delta^{-10} n^{1/2+\varepsilon+\beta}).$$

For each $v \in V$ we know that all δn triples in M_v contain two elements in one of the sets $S_1, \ldots, S_m$. Let P_v denote the set of all these pairs. That is, for each S_i, add to P_v all the pairs in S_i that are contained in a triple from M_v. We fix some arbitrary way to associate each pair in P_v with a *single* set S_i. (If this pair is in more than one set S_i just pick one arbitrarily.)

The properties of the sets $P_v, v \in V$ are summarized in the following claim.

Claim 9.3. *Each P_v is a matching of at least δn pairs, each pair $(u,w) \in P_v$ is associated with a unique S_i such that $u,w \in S_i$ and there exists a triple in M_v containing both u and w.*

Recalling the high level overview given in Section 3, we now define a way to sample a random subset in V that, using the clustering assumption, "hits" many pairs in many triples. This sampling process is given below.

The distribution μ. We denote by $\mathrm{neg}(n)$ any function of n that is less than $\exp(-n^{\alpha})$ for some constant $\alpha > 0$ and all sufficiently large n. We use the notation $A \sim \mu$ to mean "the random variable A is sampled according to the distribution μ."

We now define a distribution μ on subsets of V. To pick a set $A \sim \mu$ we first pick an index $i \in [m]$ uniformly at random and then pick $A \subset S_i$ to contain each element of S_i independently with probability $n^{-1/4+\varepsilon}$. If S_i happens to be empty, we let A be the empty set. It will be convenient to treat μ also as a distribution on pairs of the form (A,i) with $A \subset V$ and $i \in [m]$ so, we will sometimes write $(A,i) \sim \mu$ to denote that i is the random index chosen in the sampling process of A and, other times just write $A \sim \mu$.

Claim 9.4. *Let $A \sim \mu$. Then*

$$\Pr[|A| \geq n^{1/4+3\varepsilon}] \leq \mathrm{neg}(n)\,.$$

Proof. Conditioning on the choice of the set S_i, the expectation of $|A|$ is at most

$$|S_i| \cdot n^{-1/4+\varepsilon} \leq O(\delta^6 n^{1/4+2\varepsilon}) < n^{1/4+3\varepsilon}/100\,.$$

Thus, by a Chernoff bound, the probability that the size of A exceeds $n^{1/4+3\varepsilon}$ is at most $\mathrm{neg}(n)$. Taking a union bound over the m possible choices of S_i the probability is still $\mathrm{neg}(n)$. □

Observation 9.5. *We can define a new distribution μ' that samples A according to μ until it gets a set A of size at most $n^{1/4+3\varepsilon}$. By the claim, the statistical distance between μ and μ' is at most* $\mathrm{neg}(n)$. *Hence, as long as we can tolerate a* $\mathrm{neg}(n)$ *error in our probabilities, we can switch between μ and μ' as needed.*

The functions $f_{A,i}(v)$. For each set $A \subset S_i$ we define a partial function $f_{A,i} : V \to V$. The value $f_{A,i}(v)$ is defined as follows: Consider the pairs in P_v that are associated with S_i. If one of these pairs is contained in A then $f_{A,i}(v)$ is defined to be the third element of the triple of M_v associated with that pair. More formally, if there is a pair $u,w \in S_i$ so that a triple (u,w,z) is in M_v then we define $f_{A,i}(v) = z$. If there is more than one such pair, we pick one arbitrarily, for instance the first one in some fixed order. If there is no such pair, we let $f_{A,i}(v) = \perp$ (undefined).

We use the notation $x \sim y$, with $x,y \in \mathbb{F}^d$, to denote that x is a constant multiple of y and y is a constant multiple of x. That is, either they are both zero, or they are both non-zero multiples of each other. Notice that the relation $\sim$ is an equivalence relation.

Claim 9.6. *Let $i \in [m], A \subset S_i$ and let $f_{A,i}$ be defined as above. If $\mathcal{L} : \mathbb{F}^d \to \mathbb{F}^d$ is any linear map sending A to zero, then $\mathcal{L}(v) \sim \mathcal{L}(f_{A,i}(v))$ for all v for which $f_{A,i}(v) \neq \perp$.*

Proof. If $f_{A,i}(v) \neq \perp$ then there is a triple $(x,y,z) \in M_v$ with $x,y \in A$ and $f_{A,i}(v) = z$. Since $v \in \mathsf{span}\{x,y,z\}$ we get that $\mathcal{L}(v) \in \mathsf{span}\{\mathcal{L}(x),\mathcal{L}(y),\mathcal{L}(z)\} = \mathsf{span}\{\mathcal{L}(f_{A,i}(v))\}$. Similarly, since we are assuming that v is not in the span of x,y (since the matchings M_v are in regular form), z is in the span of v,x,y and so $\mathcal{L}(z) \in \mathsf{span}\{\mathcal{L}(v)\}$. □

Probability bounds. The following three claims give bounds on certain probabilities involving the functions $f_{A,i}$, when $(A,i) \sim \mu'$.

Claim 9.7. *Let $(A,i) \sim \mu'$ and let $v \in V$. Then,* $\Pr[f_{A,i}(v) \neq \perp] \geq \Omega(\delta^{17} n^{-3\varepsilon})$.

Proof. By Observation 9.5, it is enough to analyze the probability for the distribution μ. Fixing $v \in V$ we call a set S_i *heavy* if it contains at least $n^{1/2-2\varepsilon}$ pairs from P_v (recall Claim 9.3). Since we are choosing each element of S_i with probability $n^{-1/4+\varepsilon}$, the probability to "miss" a single pair from P_v is exactly $(1-n^{-1/2+2\varepsilon})$. If S_i is heavy, then (using the fact that P_v is a matching) the probability that A contains at least one of the pairs in P_v is at least

$$\Pr\left[P_v \cap \binom{A}{2} \neq \emptyset\right] \geq 1 - \left(1 - n^{-1/2+2\varepsilon}\right)^{n^{1/2-2\varepsilon}} \geq 1/2\,. \tag{9.1}$$

We now bound from below the probability that S_i is heavy. Recall that $|P_v| \geq \delta n$ and that $m \leq O(\delta^{-10} n^{1/2+\varepsilon+\beta})$. Let $m_h + m_\ell = m$ so that m_h is the number of heavy sets S_i. Since each S_i can contain at most $|S_i|/2 = O(\delta^{-6} n^{1/2+\varepsilon})$ disjoint pairs, we have that

$$\begin{aligned} \delta n &\leq m_\ell \cdot n^{1/2-2\varepsilon} + m_h \cdot O(\delta^{-6} n^{1/2+\varepsilon}) \\ &\leq O(\delta^{-10} n^{1-\varepsilon+\beta}) + m_h \cdot O(\delta^{-6} n^{1/2+\varepsilon})\,. \end{aligned}$$

This implies (since $\beta < \varepsilon/2$) that

$$m_h \geq \Omega(\delta^7 n^{1/2-\varepsilon})\,.$$

Therefore,

$$\frac{m_h}{m} \geq \Omega\left(\frac{\delta^7 n^{1/2-\varepsilon}}{\delta^{-10} n^{1/2+\varepsilon+\beta}}\right) = \Omega(\delta^{17} n^{-3\varepsilon})\,.$$

Combining the above two bounds, we get that the probability of picking a heavy cluster *and* then picking some pair in P_v is at least $\Omega(\delta^{17} n^{-3\varepsilon})$. □

Claim 9.8. *Let $(A,i) \sim \mu'$. Then, for all $v,z \in V$,*

$$\Pr[f_{A,i}(v) = z] \leq O(\delta^{-19} n^{-1+6\varepsilon})\,.$$

Proof. By Observation 9.5, it is enough to analyze the probability for the distribution μ. Suppose z appears in a triple $(u,w,z) \in M_v$ that is associated with $S_{\hat{i}}$ for some $\hat{i} \in [m]$. (If there is no such $\hat{i}$ then the probability in question is equal to zero.)By our definition of the functions $f_{A,i}$, it is only possible for $f_{A,i}(v) = z$ to hold if $i = \hat{i}$ and both u and w are chosen to be in the set $A \subset S_{\hat{i}}$. The probability to pick $i = \hat{i}$ is $1/m \leq O(\delta^{-19} n^{-1/2+3\varepsilon+\beta})$. Now, conditioned on picking this event, the probability of picking both u and w to be in A is $n^{-1/2+2\varepsilon}$. Multiplying, and using the bound $\beta < \varepsilon/2$, we get the required bound. □

Claim 9.9. *Let $(A,i) \sim \mu'$ and let $B \subset V$ be a set with $|B| \leq n^{1-10\varepsilon}$. Then, for every $v \in V$,*

$$\Pr[f_{A,i}(v) \neq \perp \wedge f_{A,i}(v) \notin B] \geq \Omega(\delta^{17} n^{-3\varepsilon}).$$

Proof. Let $p = \Pr[f_{A,i}(v) \neq \perp \wedge f_{A,i}(v) \notin B]$. Then, by Claims 9.7 and 9.8, we have

$$\begin{aligned} 1-p &= \Pr[f_{A,i}(v) = \perp \vee f_{A,i}(v) \in B] \\ &\leq \Pr[f_{A,i}(v) = \perp] + \Pr[f_{A,i}(v) \in B] \\ &\leq 1 - \Omega(\delta^{17} n^{-3\varepsilon}) + |B| \cdot O(\delta^{-19} n^{-1+6\varepsilon}) \\ &\leq 1 - \Omega(\delta^{17} n^{-3\varepsilon}) + O(\delta^{-19} n^{-4\varepsilon}). \end{aligned}$$

Rearranging, and using the fact that $n \geq (1/\delta)^{\omega(1)}$, we get that $p \geq \Omega(\delta^{17} n^{-3\varepsilon})$. □

The set U. To define the set U required in Lemma 9.2, we proceed as follows. Let r be an integer to be determined later, and pick r sets $A_1, \ldots, A_r \subset V$ and r indices $i_1, \ldots, i_r \in [m]$ so that each (A_j, i_j) is sampled independently according to the distribution μ'. Let $U = \bigcup_{j=1}^{r} A_j$. Let $f_1 = f_{A_1,i_1}, \ldots, f_r = f_{A_r,i_r}$ be the corresponding (partial) functions on V. Our goal is to show that, with probability greater than zero, setting U to zero by a linear map, reduces the dimension of V to $n^{10\varepsilon}$.

We begin by defining a sequence of undirected graphs $H_0, H_1, \ldots, H_r$ on vertex set V which will depend on the choice of the sets $A_1, \ldots, A_r$. The first graph H_0 is the empty graph (containing no edges). We define H_j inductively by adding to H_{j-1} all edges of the form $(v, f_j(v))$ over all $v \in V$. For $j = 1, \ldots, r$, let k_j denote the number of connected components of H_j.

Claim 9.10. *If $\mathcal{L} : \mathbb{F}^d \to \mathbb{F}^d$ is any linear map sending U to zero, then $\mathsf{span}\{\mathcal{L}(V)\}$ has dimension at most k_r.*

Proof. This is an easy corollary of Claim 9.6. If $\mathcal{L}(U) = 0$ then, for every edge (x,y) in H_r, we have $\mathcal{L}(x) \sim \mathcal{L}(y)$. Since the relation $\sim$ is transitive, each connected component is contained in a one dimensional subspace after applying $\mathcal{L}$. □

Let k'_j denote the number of connected components of H_j of size at most $n^{1-10\varepsilon}$. Call these the "small" components of H_j. The next claim bounds the expectation of k'_j.

Claim 9.11. *Let $1 \leq j \leq r$. Then,*

$$\mathbb{E}[k'_j] \leq k'_{j-1}(1 - \Omega(\delta^{17} n^{-3\varepsilon})).$$

Proof. Let $s = k'_{j-1}$ and let $K_1, \ldots, K_s$ be the small components of H_{j-1}. Pick representatives $u_i \in K_i$ in each of the components. For each $i = 1, \ldots, s$, let X_i be an indicator variable so that $X_i = 1$ if $f_j(u_i) \in V \setminus K_i$ (that is, $f_j(u_i)$ is defined and does not belong to K_i) and $X_i = 0$ otherwise (if either $f_j(u_i) = \perp$ or if it is defined but in K_i). By Claim 9.9, we have that

$$\mathbb{E}[X_i] = \Pr[X_i = 1] \geq \Omega(\delta^{17} n^{-3\varepsilon}).$$

Since having an edge $(u_i, f_j(u_i))$ going from u_i to some vertex outside K_i "merges" K_i with another component, we have that

$$k'_j \leq s - \frac{1}{2} \sum_{i=1}^{s} X_i.$$

Taking expectations, and using the above bound on the expectations of the X_i, we get

$$\mathbb{E}[k'_j] \leq s(1 - \Omega(\delta^{17} n^{-3\varepsilon}))$$

as was required. □

Thus, for each $j = 1, 2, \ldots, r$ there is a choice of a set $A_j \subset S_{i_j}$ such that H_j has at most $k'_{j-1}(1 - \Omega(\delta^{17} n^{-3\varepsilon}))$ small components. Taking $r = n^{4\varepsilon}$, we get that there is a choice of U for which H_r does not have *any* small components. Since the number of large components is at most $n^{10\varepsilon}$, we get:

Claim 9.12. *There is a choice of U for which H_r has at most $n^{10\varepsilon}$ connected components.*

To conclude, we observe that, since we are using the modified distribution μ', we have

$$|U| \leq r \cdot n^{1/4+3\varepsilon} \leq n^{1/4+7\varepsilon}$$

and, using Claim 9.10, we have that, setting U to zero by a linear map, reduces the dimension of V to at most $n^{10\varepsilon}$. This completes the proof of Lemma 9.2.

10 Putting it all together: proof of Theorem 2.3

We will first prove that any $(3, \delta)$-LCC over $\mathbb{R}$ contains a large subset of small dimension. Later we will iterate this to get a global dimension bound.

Lemma 10.1. *Suppose $n > (1/\delta)^{\omega(1)}$ and let $0 < \varepsilon < 1/50$. Let $V = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$ be a $(3, \delta)$-LCC. Then, there exists a subset $U \subset V$ of size at least*

$$|U| \geq (\delta^3/300)n$$

and dimension at most

$$\dim(U) \leq \max\{8\delta^6 d, n^{1/2-\varepsilon/16}\}.$$

Proof. We will prove the lemma by first applying Lemma 8.2 to show that V has a large sub-LCC V' in which the triples cluster. Then, we will apply Lemma 9.1 to show that V' has a large low-rank sublist. The details follow.

Set $\beta_1 = \varepsilon/4$ and apply Lemma 8.2 with $\beta = \beta_1$. To apply the lemma we require that V does not contain a subset U of size $(\delta^2/288)n$ and rank at most $\max\{8\delta^6 d, n^{1/2-\beta_1/4}\} = \max\{8\delta^6 d, n^{1/2-\varepsilon/16}\}$. If this is the case, then our proof is done and there is no need to continue.

Having applied Lemma 8.2, we obtain a $(3,\delta')$-LCC $V' \subset V$ with $n' = |V'| \geq (\delta/10)n$, $d' = \dim(V') \leq d$, $\delta' \geq \delta^2/4$ and sets $S_1, \ldots, S_m$ which cluster all the triples in the matchings $M_{v'}, v' \in V'$ used to decode V' so that

$$|S_i| \leq O(n'/\delta'^6 d')$$

and

$$\Omega(\delta'^{19} d'^3/n'^{1+2\beta_1}) \leq m \leq O(n'^{1+2\beta_1}/\delta'^{10} d').$$

We now apply Lemma 9.1 with $\beta = 2\beta_1 < \varepsilon/2$ and the same ε to conclude that there exist a subset $V'' \subset V'$ of size

$$n'' = |V''| \geq (\delta'/2)n' \geq (\delta^2/8)(\delta/10)n \geq (\delta^3/80)n$$

and dimension

$$\dim(V'') \leq n''^{1/2-\varepsilon} \leq \max\{8\delta^6 d, n^{1/2-\varepsilon/16}\},$$

as was required. □

We now prove an amplification lemma which uses Lemma 10.1 iteratively. For this lemma we will use the following convenient notation. If $S \subset V$ is a subset of V, we denote by $\mathsf{span}_V(S) \subset V$ the subset of elements of V that are spanned by elements of S. (We think of all these as lists/multisets.)

Lemma 10.2 (Amplification lemma)**.** *Suppose $n > (1/\delta)^{\omega(1)}$ and let $0 < \varepsilon < 1/50$. Let $V = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$ be a linear $(3,\delta)$-LCC. Suppose $S \subset V$ is such that $\mathsf{span}_V(S) = S$ and $S \neq V$. Then there is a set $S \subseteq S' \subseteq V$ with $\mathsf{span}_V(S') = S'$ such that*

1. *either $S' = V$ or $|S'| \geq |S| + (\delta^4/400)n$;*
2. $\dim(S') \leq \dim(S) + \max\{\delta^6 d, n^{1/2-\varepsilon/16}\}$.

We defer the proof of the lemma to the end of this section and proceed with the proof of Theorem 2.3.

Proof of Theorem 2.3. Let $V = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$ be a linear $(3,\delta)$-LCC. We will prove the theorem with $\varepsilon = 1/1000$ We now apply Lemma 10.2 with $\varepsilon_1 = 1/51$ iteratively. Start with $S_1 = \emptyset$ and apply Lemma 10.2 repeatedly to obtain sets $S_2, S_3, \ldots$, such that for all i,

$$|S_i| \geq |S_{i-1}| + (\delta^4/400)n$$

and

$$\dim(S_i) \leq \dim(S_{i-1}) + \max\{\delta^6 d, n^{1/2-\varepsilon_1/16}\}.$$

Since the size of S_i cannot grow beyond n, the process will terminate after at most $m = \lfloor 400/\delta^4 \rfloor$ steps, yielding $S_m = V$. We then get that

$$\dim(S_m) = \dim(V) \leq (400/\delta^4)\max\{\delta^6 d, n^{1/2-\varepsilon_1/16}\} = \max\{(400\delta^2)d, (400/\delta^4)n^{1/2-\varepsilon_1/16}\}.$$

Without loss of generality, for the proof of the theorem we can assume that $\delta^2 < 1/500$. Thus it must be that

$$d = \dim(V) \leq (400/\delta^4)n^{1/2-\varepsilon_1/16} \leq n^{1/2-\varepsilon}.$$

This completes the proof of Theorem 2.3. □

10.1 Proof of Lemma 10.2

Observe that for $v \in V \setminus S$, all 3 points of any triple in M_v cannot be in S since $\mathsf{span}_V(S) = S$. Thus we may assume that $|S| \leq (1-\delta)n$, since otherwise each vector in $V \setminus S$ would be spanned by the points of S and we would be done.

Case 1: There exists a $v \in V \setminus S$ such that $\delta n/4$ of the triples in M_v have two of their points contained in S. In this case let $S' = \mathsf{span}_V(\{v\} \cup S)$. Then $|S'| \geq |S| + (\delta/4)n$, and $\dim(S') \leq \dim(S) + 1$.

If Case 1 does not hold then each $v \in V \setminus S$, M_v has $3\delta n/4$ of its triples intersecting S in either one or zero points. Let us call a point v *type-zero* if if has at least $3\delta n/8$ of its triples contained in $V \setminus S$ and *type-one* otherwise. Notice that, if v is type-one, then it must have at least $3\delta n/8$ of its triples intersecting S in exactly one point. We now separate into two additional cases.

Case 2: There are at most $\delta n/4$ type-one points. Let $V' \subset V \setminus S$ be the set of all type-zero points. Observe that, since $|S| \leq (1-\delta)n$, we have $|V'| \geq 3\delta n/4$. Also observe that the vectors in V' form a $(3, \delta/8)$-LCC since each point in V' has at least $3\delta n/8 - \delta n/4 = \delta n/8 \geq (\delta/8)|V'|$ triples in its matching contained in V'. Using Lemma 10.1 on V' we conclude that there is a subset $U \subset V'$ of size

$$|U| \geq (\delta^3/300)|V'| \geq (\delta^4/400)n$$

and dimension

$$\dim(U) \leq \max\{8(\delta/8)^6 d', |V'|^{1/2-\varepsilon/16}\} \leq \max\{\delta^6 d, n^{1/2-\varepsilon/16}\}.$$

Setting $S' = S \cup U$ we are done.

Case 3: There are at least $\delta n/4$ type-one points. In this case, there are $\delta n/4$ points v in $V \setminus S$, each having at least $3\delta n/8$ of the triples in M_v intersecting S in exactly one point. Let A be a linear transformation whose kernel equals $\mathsf{span}(S)$. After applying A to $V \setminus S$ we obtain a $(2, 3\delta/4)$ LDC decoding the $\delta n/4$ type-one points. Thus the $\delta n/4$ points (after we apply the mapping A to them) must span at most $\mathsf{poly}(1/\delta)\log n \leq \max\{\delta^6 d, n^{1/2-\varepsilon/16}\}$ dimensions by Theorem 4.5. Thus, adding them to S will increase the dimension of its span by at most this number. This completes the proof also in this case. □

References

[1] SANJEEV ARORA, CARSTEN LUND, RAJEEV MOTWANI, MADHU SUDAN, AND MARIO SZEGEDY: Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, 1998. Preliminary versions in FOCS'92 and ECCC. [doi:10.1145/278298.278306] 2

[2] SANJEEV ARORA AND SHMUEL SAFRA: Probabilistic checking of proofs: A new characterization of NP. *J. ACM*, 45(1):70–122, 1998. Preliminary version in FOCS'92. [doi:10.1145/273865.273901] 2

[3] LÁSZLÓ BABAI, LANCE FORTNOW, AND CARSTEN LUND: Non-deterministic exponential time has two-prover interactive protocols. *Comput. Complexity*, 1(1):3–40, 1991. Preliminary version in FOCS'90. [doi:10.1007/BF01200056] 2

[4] LÁSZLÓ BABAI, LANCE FORTNOW, NOAM NISAN, AND AVI WIGDERSON: BPP has subexponential time simulations unless EXPTIME has publishable proofs. *Comput. Complexity*, 3(4):307–318, 1993. Preliminary version in SCT'91. [doi:10.1007/BF01275486] 2

[5] BOAZ BARAK, ZEEV DVIR, AMIR YEHUDAYOFF, AND AVI WIGDERSON: Rank bounds for design matrices with applications to combinatorial geometry and locally correctable codes. In *Proc. 43rd STOC*, pp. 519–528. ACM Press, 2011. [doi:10.1145/1993636.1993705, arXiv:1009.4375] 2, 3, 6, 8, 9

[6] FRANCK BARTHE: On a reverse form of the Brascamp-Lieb inequality. *Inventiones Math.*, 134(2):335–361, 1998. [doi:10.1007/s002220050267, arXiv:math/9705210] 11, 12, 13

[7] DONALD BEAVER AND JOAN FEIGENBAUM: Hiding instances in multioracle queries. In *Proc. 7th Symp. Theoretical Aspects of Comp. Sci. (STACS'90)*, volume 415 of *LNCS*, pp. 37–48. Springer, 1990. [doi:10.1007/3-540-52282-4_30] 2

[8] ARNAB BHATTACHARYYA, ZEEV DVIR, AMIR SHPILKA, AND SHUBHANGI SARAF: Tight lower bounds for 2-query LCCs over finite fields. *Combinatorica*, 36(1):1–36, 2016. Preliminary version in FOCS'11. [doi:10.1007/s00493-015-3024-z] 3, 6, 9

[9] MANUEL BLUM AND SAMPATH KANNAN: Designing programs that check their work. *J. ACM*, 42(1):269–291, 1995. Preliminary version in STOC'89. [doi:10.1145/200836.200880] 2

[10] MANUEL BLUM, MICHAEL LUBY, AND RONITT RUBINFELD: Self-testing/correcting with applications to numerical problems. *J. Comput. Sys. Sci.*, 47(3):549–595, 1993. Preliminary version in STOC'90. [doi:10.1016/0022-0000(93)90044-W] 2

[11] BENNY CHOR, EYAL KUSHILEVITZ, ODED GOLDREICH, AND MADHU SUDAN: Private information retrieval. *J. ACM*, 45(6):965–981, 1998. Preliminary version in FOCS'95. [doi:10.1145/293347.293350] 2

[12] ZEEV DVIR: On matrix rigidity and locally self-correctable codes. *Comput. Complexity*, 20(2):367–388, 2011. Preliminary version in CCC'10. [doi:10.1007/s00037-011-0009-1] 2

[13] ZEEV DVIR, PARIKSHIT GOPALAN, AND SERGEY YEKHANIN: Matching vector codes. *SIAM J. Comput.*, 40(4):1154–1178, 2011. Preliminary version in FOCS'10. [doi:10.1137/100804322] 2

[14] ZEEV DVIR, SHUBHANGI SARAF, AND AVI WIGDERSON: Breaking the quadratic barrier for 3-LCC's over the reals. In *Proc. 46th STOC*, pp. 784–793. ACM Press, 2014. [doi:10.1145/2591796.2591818, arXiv:1311.5102] 1

[15] ZEEV DVIR, SHUBHANGI SARAF, AND AVI WIGDERSON: Improved rank bounds for design matrices and a new proof of Kelly's theorem. *Forum of Math., Sigma*, 2(4), 2014. [doi:10.1017/fms.2014.2, arXiv:1211.0330] 2, 3, 8

[16] ZEEV DVIR AND AMIR SHPILKA: Locally decodable codes with 2 queries and polynomial identity testing for depth 3 circuits. *SIAM J. Comput.*, 36(5):1404–1434, 2007. Preliminary version in STOC'05. [doi:10.1137/05063605X] 2, 3, 11

[17] KLIM EFREMENKO: 3-query locally decodable codes of subexponential length. *SIAM J. Comput.*, 41(6):1694–1703, 2012. Preliminary version in STOC'09. [doi:10.1137/090772721] 2

[18] ODED GOLDREICH, HOWARD J. KARLOFF, LEONARD J. SCHULMAN, AND LUCA TREVISAN: Lower bounds for linear locally decodable codes and private information retrieval. *Comput. Complexity*, 15(3):263–296, 2006. Preliminary version in CCC'02. [doi:10.1007/s00037-006-0216-3] 3

[19] JONATHAN KATZ AND LUCA TREVISAN: On the efficiency of local decoding procedures for error-correcting codes. In *Proc. 32nd STOC*, pp. 80–86. ACM Press, 2000. [doi:10.1145/335305.335315] 2, 3

[20] NEERAJ KAYAL AND SHUBHANGI SARAF: Blackbox polynomial identity testing for depth 3 circuits. In *Proc. 50th FOCS*, pp. 198–207. IEEE Comp. Soc. Press, 2009. [doi:10.1109/FOCS.2009.67] 2

[21] IOANNIS KERENIDIS AND RONALD DE WOLF: Exponential lower bound for 2-query locally decodable codes via a quantum argument. *J. Comput. Sys. Sci.*, 69(3):395–420, 2004. Preliminary version in STOC'03. [doi:10.1016/j.jcss.2004.04.007, arXiv:quant-ph/0208062] 3

[22] SWASTIK KOPPARTY: List-decoding multiplicity codes. *Theory of Computing*, 11(5):149–182, 2015. [doi:10.4086/toc.2015.v011a005] 2

[23] SWASTIK KOPPARTY, SHUBHANGI SARAF, AND SERGEY YEKHANIN: High-rate codes with sublinear-time decoding. *J. ACM*, 61(5):28:1–28:20, 2014. Preliminary version in STOC'11. [doi:10.1145/2629416] 2

[24] PETER D. LAX: *Linear Algebra and Its Applications*. Wiley, 2007. 13

[25] RICHARD J. LIPTON: Efficient checking of computations. In *Proc. 7th Symp. Theoretical Aspects of Comp. Sci. (STACS'90)*, volume 415 of *LNCS*, pp. 207–215. Springer, 1990. [doi:10.1007/3-540-52282-4_44] 2

[26] CARSTEN LUND, LANCE FORTNOW, HOWARD J. KARLOFF, AND NOAM NISAN: Algebraic methods for interactive proof systems. *J. ACM*, 39(4):859–868, 1992. Preliminary version in FOCS'90. [doi:10.1145/146585.146605] 2

[27] RONITT RUBINFELD AND MADHU SUDAN: Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996. [doi:10.1137/S0097539793255151] 2

[28] ADI SHAMIR: IP = PSPACE. *J. ACM*, 39(4):869–877, 1992. Preliminary version in FOCS'90. [doi:10.1145/146585.146609] 2

[29] LESLIE G. VALIANT: Graph-theoretic arguments in low-level complexity. In *Proc. 6th Symp. Math. Found. Comput. Sci. (MFCS'77)*, volume 53 of *LNCS*, pp. 162–176. Springer, 1977. [doi:10.1007/3-540-08353-7_135] 2

[30] DAVID P. WOODRUFF: New lower bounds for general locally decodable codes. *Electron. Colloq. on Comput. Complexity (ECCC)*, 14(6), 2007. Available at ECCC. 3

[31] DAVID P. WOODRUFF: A quadratic lower bound for three-query linear locally decodable codes over any field. *J. Comput. Sci. Techn.*, 27(4):678–686, 2012. Preliminary version in RANDOM'10. [doi:10.1007/s11390-012-1254-8] 3

[32] SERGEY YEKHANIN: Towards 3-query locally decodable codes of subexponential length. *J. ACM*, 55(1):1:1–1:16, 2008. Preliminary version in STOC'07. [doi:10.1145/1326554.1326555] 2

AUTHORS

Zeev Dvir
Associate professor
Department of Computer Science and Department of Mathematics
Princeton University
zdvir@princeton.edu
http://www.cs.princeton.edu/~zdvir/

Shubhangi Saraf
Assistant professor
Department of Computer Science and Department of Mathematics
Rutgers University
shubhangi.saraf@rutgers.edu
https://www.math.rutgers.edu/~ss1984/

Avi Wigderson
Professor
School of Mathematics
Institute for Advanced Study
avi@ias.edu
http://www.math.ias.edu/~avi

ABOUT THE AUTHORS

ZEEV DVIR was born in Jerusalem, Israel. He received his Ph. D. from the Weizmann Institute in Israel in 2008. His advisors were Ran Raz and Amir Shpilka. He has a broad interest in theoretical computer science and mathematics and especially in computational complexity, pseudorandomness, coding theory and discrete mathematics.

SHUBHANGI SARAF grew up in Pune, India. She received her Ph. D. from the Massachusetts Institute of Technology under the guidance of Madhu Sudan. Details about Shubhangi's early career can be found in her bio sketch in Volume 13, Article 6 of *Theory of Computing*.

Shubhangi is broadly interested in complexity theory, coding theory and pseudorandomness. Recently she has been captivated by questions related to understanding the power and limitations of algebraic computation, as well as to understanding the potential of "locality" in algorithms for codes. In her spare time Shubhangi enjoys reading, cooking, long walks, and exploring cafés and restaurants. Her little toddler is a constant source of joy and amazement, and also makes sure there isn't much time to spare.

AVI WIGDERSON was born in Haifa, Israel in 1956, and received his Ph. D. in 1983 at Princeton University under Dick Lipton. He enjoys and is fascinated with studying the power and limits of efficient computation, and the remarkable impact of this field on understanding our world. Avi's other major source of fascination and joy are his three kids, Eyal, Einat, and Yuval.