Tight lower bounds for 2-query LCCs over finite fields

Arnab Bhattacharyya
Dept. of Computer Science
Princeton University
Princeton, NJ
Email: arnabb@princeton.edu

Zeev Dvir
Dept. of Computer Science
Princeton University
Princeton, NJ
Email: zdvir@princeton.edu

Shubhangi Saraf
School of Mathematics
Institute for Advanced Study
Princeton, NJ
Email: shubhangi@ias.edu

Amir Shpilka
Faculty of Computer Science
Technion — Israel Institute of Technology
Haifa, Israel
Email: shpilka@cs.technion.ac.il

Abstract— A Locally Correctable Code (LCC) is an error correcting code that has a probabilistic self-correcting algorithm that, with high probability, can correct any coordinate of the codeword by looking at only a few other coordinates, even if a fraction δ of the coordinates are corrupted. LCCs are a stronger form of LDCs (Locally Decodable Codes) which have received a lot of attention recently due to their many applications and surprising constructions.

In this work we show a separation between 2-query LDCs and LCCs over finite fields of prime order. Specifically, we prove a lower bound of the form $p^{\Omega(\delta d)}$ on the length of linear 2-query LCCs over $\mathbb{F}_p$, that encode messages of length d. Our bound improves over the known bound of $2^{\Omega(\delta d)}$ [9], [12], [8] which is tight for LDCs. Our proof makes use of tools from additive combinatorics which have played an important role in several recent results in theoretical computer science.

Corollaries of our main theorem are new incidence geometry results over finite fields. The first is an improvement to the Sylvester-Gallai theorem over finite fields [14] and the second is a new analog of Beck's theorem over finite fields.

Keywords-locally decodable codes; sylvester-gallai theorem; additive combinatorics

1. Introduction

Locally Correctable Codes (LCCs) are special families of error correcting codes (ECCs) which possess an additional structure. Besides being able to recover a message from its noisy transmission (the original purpose of ECCs, as defined by Shannon [15]), these codes enable the receiver to recover any single coordinate of the codeword from a 'local' sample of the other, possibly corrupted, coordinates. The local correction is guaranteed to work with high probability as long as the number of errors is not too large. Roughly speaking, a linear q-query locally correctable code ((q,δ)-LCC for short) over a field $\mathbb{F}_p$ is a subspace $C \subseteq \mathbb{F}_p^n$ such that, given an element $\tilde{y}$ that disagrees with some $y \in C$ in at most δn coordinates and an index $i \in [n]$, one can recover y_i with, say, probability 0.9, by reading at most q coordinates of $\tilde{y}$. In this setting, the 'message length' is the dimension of C, or $d = \log_p(|C|)$.

The notion of LCCs was preceded in the literature by the weaker notion of Locally Decodable Codes (LDCs) in which one has the seemingly weaker property that message symbols (as opposed to codeword symbols) are to be 'locally decoded'. In fact, for linear codes, which are our main interest, LCCs are a subfamily of LDCs (since every linear code can be assumed to be systematic and therefore local correction implies local decoding). Both LDCs and LCCs have many applications in theoretical computer science. See [16] for a survey of these codes and their uses.

The main question with respect to LCCs (or LDCs) is how good can they be. That is, what limitations can we prove on their encoding length, as a function of the message length, the number of queries and the amount of error the decoder can tolerate. Our knowledge in this area is very limited, and considerable gaps between lower and upper bounds exist when the number of queries is larger than two.

In this work we focus on the simplest question of this form. Assuming that the message length is d and the underlying field is $\mathbb{F}_p$, "*What is the minimal encoding length n for which we can recover any symbol of any codeword by making just 2 queries, assuming that less than δn coordinates were corrupted?*"

One motivation for studying this question comes from the desire to better understand the relation between LDCs and LCCs and explain the lack of constructions for LCCs. Although it may seem surprising, the question of proving a lower bound for LCCs with 2 queries is a fundamental problem that lies in the core of many questions in geometry, additive combinatorics and more. As we shall see, similar to some of the connections made in [3], the question that we study here is closely related to questions such as: generalizations of the famous Sylvester-Gallai theorem; extensions of Beck's theorem; proving lower bounds on the rank of matrices that satisfy certain 'design' like properties. Our techniques also highlight a close connection of LCCs to problems in additive combinatorics. We later expand on each of the problems and state our contributions.

Our main theorem is a tight lower bound for linear LCCs over $\mathbb{F}_p$, improving the exponential lower bound, that was proved in [9], [12], [8] for LDCs, $n > 2^{\Omega(\delta d)}$, where d is the message length, to $n > p^{\Omega(\delta d)}$ for all constants p and δ. A formal statement is given in the section below.

1.1. The Main Theorem

Denote by $\mathbb{F}_p$ the field of residues modulo a prime number p. When working with 2-query linear LCCs, it will be convenient to adopt a 'geometrical' way of looking at those codes and speak of their dimension instead of message length. Note that, for such codes, it is well known that the decoding can be made linear as well without loss of generality while only losing a constant factor (depending on the number of queries) in the error (see [3]).

Definition 1.1 (Linear 2-query LCC)**.** *Let $V = (v_1, \dots, v_n) \in (\mathbb{F}_p^d)^n$ be a list of n vectors (possibly with repetitions) in $\mathbb{F}_p^n$. We say that V is a $(2, \delta)$*-LCC *(Locally Correctable Code) if for every $i \in [n]$ and every subset $S \subseteq [n]$ of size at most δn, there exist a pair of indices $j, j' \in [n] \setminus S$ such that $v_i \in \mathrm{span}\{v_j, v_{j'}\}$. We let $\dim(V)$, the* dimension *of V, denote the dimension of the span of the vectors $v_1, \dots, v_n$ inside $\mathbb{F}_p^d$.*

To see the connection to the (sketchy) definition given in the previous section, we note that C is the subspace that is spanned by the rows of the $d \times n$ matrix G whose columns are $(v_1, \dots, v_n)$. One can think of encoding a message of length d, $\bar{a} = (a_1, \dots, a_d)$, as $\mathrm{Enc}(\bar{a}) = \bar{a} \cdot G$. We also note that in several previous works (e.g. in [7]), LCCs are defined by means of their dual matrix but, for our purposes, this (equivalent) definition, in terms of the generating matrix, will be more convenient.

Theorem 1 (Main Theorem)**.** *There exist universal constants $c_1, c_2 > 0$ such that for every $\epsilon > 0$ and every prime p, the following holds. Let $V = (v_1, \dots, v_n) \in (\mathbb{F}_p^d)^n$ be a $(2, \delta)$-LCC. Then*

$$\dim(V) \leq c_1(p/\epsilon\delta)^{c_2} + ((2+\epsilon)/\delta) \cdot \log_p(n).$$

In particular, if we wish to linearly encode a message of length d using a 2-query LCC, then we must have $n = \Omega_{p,\delta,\epsilon}\left(p^{\frac{\delta}{2+\epsilon}d}\right)$.

1.2. Previous work

As mentioned above, each LCC is also an LDC, so lower bounds for LDCs give lower bounds for LCCs. Exponential lower bounds (i.e., $n \geq \exp(d)$) for LDC's were proven for two-query codes (also for non-linear codes) in [9], [12], [8]. These bounds are tight since the Hadamard code achieves $n = 2^d$ and is locally decodable for constant δ. We remind the reader that the Hadamard code is a linear code over $\mathbb{F}_2$ which takes a message $x \in \mathbb{F}_2^d$ and encodes it as a codeword of length 2^d given by

$$H(x) = (\langle a, x\rangle)_{a \in \mathbb{F}_2^d}.$$

This gives a linear 2-query LDC with constant δ since, to recover x_i, we can query $\langle a, x\rangle$ and $\langle a + e_i, x\rangle$ for random $a \in \mathbb{F}_2^d$ (where e_i denote the i'th unit vector in the standard basis). This is also a linear LCC over $\mathbb{F}_2$ since any coordinate $\langle a, x\rangle$ can also be recovered from two random positions in a similar way.

When trying to generalize the Hadamard code construction to fields $\mathbb{F}_p$ with $p > 2$ a prime number, we are faced with the following situation. To get a LDC, we can use the exact same construction described above, where we replace $\mathbb{F}_2^n$ with the set $\{0,1\}^n \subset \mathbb{F}_p^n$. One can check that decoding $x_i \in \mathbb{F}_p$ is still possible using two random queries as above. If we are interested in LCCs, however, things are much worse. The best construction we can get is essentially $C = \mathbb{F}_p^n$. That is, we encode a message using all vectors in $\mathbb{F}_p^n$. The dependence on the field size is more dramatic if we consider LCCs over fields over characteristic zero. In [3], Barak et al. proved that the message length cannot be larger than $O(1/\delta^9)$. In particular, larger messages cannot be encoded by LCCs. This shows a considerable difference between LDCs and LCCs over characteristic zero fields. However, prior to this work, no separation of LCCs and LDCs, over small finite fields, was known. Theorem 1 gives a tight lower bound for linear LCCs with 2 queries over $\mathbb{F}_p$, thus providing a separation between 2-query LCCs and LDCs, over finite fields (other than $\mathbb{F}_2$).

1.3. Incidence Geometry over Finite Fields

One natural way of viewing linear LCCs is as point configurations with certain algebraic restrictions. This is the point of view we chose to adapt in Definition 1.1, where the code was presented in the form of a list of vectors $(v_1, \dots, v_n) \in (\mathbb{F}_p^d)^n$ satisfying certain conditions on the spans of pairs of vectors. In [3] it was shown that bounds on 2-query LCCs are actually generalizations of the well-known *Sylvester-Gallai Theorem* from combinatorial geometry. Perhaps surprisingly, this theorem and its generalizations for finite fields, have recently found applications in algorithms for polynomial identity testing of depth-3 arithmetic circuits [11], [14]. The simplest form of this theorem is as follows.

Theorem 1.2 (Sylvester-Gallai theorem)**.** *If n distinct points in $\mathbb{R}^d$ are not collinear, then there exists a line that passes through exactly two of them.*

For a full discussion on the connection between LCCs and this theorem, we refer the reader to [3]. Informally, the conditions of the form $v_i \in \mathrm{span}\{v_j, v_{j'}\}$, given in Definition 1.1, correspond to the three points $v_i, v_j, v_{j'} \in \mathbb{F}_p^d$ being collinear (one has to move to projective space to see this). Thus, a 2-query linear LCC is a configuration of points with 'many' collinear triples, satisfying some combinatorial condition depending on the parameter δ. The Sylvester-Gallai theorem can be stated as saying that, if in a configuration of points, every pair of points defines a line which contains a third point, then the points span a subspace of dimension 1. Stated this way, the connection to our main theorem is clear. Both results translate information about 'dependent' triples into global bounds on the dimension of the entire set. We now give a Corollary of our main theorem, stated in the setting of the SG theorem.

Corollary 1.3 (Sylvester-Gallai for Finite Fields)**.** *Let $V = \{v_1, \ldots, v_n\} \subseteq \mathbb{F}_p^d$ be a set of n vectors, no two of which are linearly dependent. Suppose that for every $i, j \in [n]$, there exists $k \in [n]$ such that v_i, v_j, v_k are linearly dependent. Then, for every $\epsilon > 0$,*

$$\dim(V) \le \text{poly}(p/\epsilon) + (4 + \epsilon)\log_p n.$$

Previously, the best upper bound on $\dim(V)$ was $18\log_2 n = (18\log_2 p)\cdot\log_p n$, due to Saxena and Seshadhri [14]. Note that the set of points $V = \mathbb{F}_p^d$ shows that $\dim(V) \ge \log_p n$ is possible in Corollary 1.3.

Another corollary of our main theorem is a finite field analog of *Beck's Theorem* [5]. Over the reals, Beck's Theorem states that there exist positive integers α, β such that for any n points lying in the real plane, if there are at most αn^2 lines incident to at least two points, then at least βn points are collinear (i.e. belong to an affine subspace of dimension 1). Our analog below shows that, over finite fields, one can find (under the same assumption) a large subset that lies on a 'low dimension' subspace (instead of on a line).

Corollary 1.4 (Analog of Beck's Theorem for Finite Fields)**.** *Let $V = \{v_1, \ldots, v_n\}$ be a set of n vectors in $\mathbb{F}_p^d$, no two of which are linearly dependent. If the number of lines incident to at least two points of V is at most αn^2 for $\alpha < 1/64$, then there exists $V' \subseteq V$ such that $|V'| \ge |V|/2$ and for every $\epsilon > 0$,*

$$\dim(V') \le \text{poly}(p/\epsilon\delta) + ((2 + \epsilon)/\delta) \cdot \log_p n$$

where $\delta = 1 - 8\sqrt{\alpha}$.

As before, it is not hard to see that $\dim(V') \ge \log_p n$ is possible in Corollary 1.4.

Both corollaries 1.3 and 1.4 require less machinery than the proof of our main result, Theorem 1, and can be obtained in a more direct fashion by applying the same tools from additive combinatorics used in the proof of Theorem 1. The reason is that in Theorem 1, the points $v_1, \ldots, v_n$ are not assumed to be distinct whereas in the corollaries of this section, they are. The non-distinctness makes the argument for Theorem 1 much more elaborate, as we describe later. The proofs of both corollaries are omitted due to the page limit and can be found in the full version of the paper [6].

1.4. A Rank Bound for Design Matrices over Finite Fields

The connection between combinatorial properties of matrices, such as the zero/nonzero pattern of the matrix entries, and their algebraic properties, such as their rank, is a very interesting and important topic in the context of theoretical computer science. For instance, one can hope that such understanding could lead to explicit constructions of rigid matrices [3], [7]. An example of the usefulness of such bounds is demonstrated by the work of Alon [1], that proved lower bound on the ranks of *perturbed identity matrices*. That is, matrices in which all diagonal entries are significantly larger in magnitude than all other entries. Alon showed how to use this rank bound to obtain interesting results in geometry, coding theory and more. In a similar fashion, the recent work [3], that gave a lower bound on the rank of *design matrices* over the real numbers, had interesting applications in geometry (and of course was used to obtain lower bounds on LCCs over the reals). Roughly speaking, design matrices have restrictions on the number of nonzero entries per row, on the number of nonzero entries per column and on the size of pairwise intersections of sets of nonzero entries of columns. The connection between design matrices and LCCs was first observed in [4]. Specifically, [4] showed that lower bounds on LCCs are tightly connected to the problem of determining the minimum rank certain design matrices.

To explain the connection we start with a formal definition of this family of matrices.

Definition 1.5 (Design matrix)**.** *Let A be an $m \times n$ matrix over some field. For $i \in [m]$ let $R_i \subset [n]$ denote the set of indices of all non-zero entries in the i'th row of A. Similarly, let $C_j \subset [m]$, $j \in [n]$, denote the set of non-zero indices in the j'th column. We say that A is a (q, k, t)-*design matrix *if*

1) *For all $i \in [m]$, $|R_i| \le q$.*
2) *For all $j \in [n]$, $|C_j| \ge k$.*
3) *For all $j_1 \ne j_2 \in [n]$, $|C_{j_1} \cap C_{j_2}| \le t$.*

The following simple claim shows the connection between these matrices and LCCs. The claim holds for all values of q but we state it for $q = 3$ since we only defined 2-query LCCs. We omit the (simple) proof and refer the reader to either [4] or [3] for more details.

Claim 1.6. *Let A be a $(3, k, t)$-design matrix with m rows and n columns over a field $\mathbb{F}$. Suppose* rank$(A) \le n - d$. *Then there exists a linear $(2, \delta)$-LCC $V = (v_1, \ldots, v_n) \in \mathbb{F}^d$ with dimension d, where $\delta = \frac{k}{2nt}$.*

Hence, we can use Theorem 1 to obtain the following corollary.

Corollary 1.7 (Rank bound for design matrices)**.** *Let $\alpha > 0$ and let A be a $(3, \alpha n, t)$-design matrix with m rows and n*

columns over a field $\mathbb{F}_p$, p prime. Then, for every $\epsilon > 0$,

$$\operatorname{rank}(A) > n - \operatorname{poly}\left(\frac{pt}{\alpha\epsilon}\right) - \frac{(4+\epsilon)t}{\alpha}\log_p(n).$$

It is an interesting open problem to generalize this bound to matrices with $q > 3$. This will not immediately imply a bound on LCCs with more than 2-queries, but will be, in our opinion, a big step towards this goal.

1.5. Organization

In Section 2 we give a high level view of the proof and the techniques used. Section 3 contains some notations and basic facts from additive combinatorics. In Section 4 we give the proof of our main result, Theorem 1 based on four auxiliary lemmas. Three of these lemmas are proven in Sections 5-7. The proof of the fourth lemma, Lemma 4.3, is omitted due to the page limit and can be found in the full version of the paper [6].

2. Overview of the proof

To describe the basic idea behind our proof, we first explain how to obtain a lower bound in the case that the LCC does not have repeated coordinates. Namely, that any two coordinates correspond to linearly independent vectors in $\mathbb{F}_p^d$. Although this may seem a bit odd, many of the technical difficulties in proving Theorem 1 stem from such possible repetitions. As we shall soon see, the proof for the case of no repetitions uses a theorem of Ruzsa from additive combinatorics concerning "approximate vector spaces". The general case follows by proving a distributional version of this theorem and involves a careful combinatorial analysis.

The difficulty in handling repeated coordinates was already noticed in [3], where analogous results were proven over the reals. The way we handle repetitions is similar in spirit to the methods of [3] but requires several new ideas. In particular, we make heavy use of the fact that the field is 'not too large' which enables us to assume that the decoding is always in the form of summing two coordinates (without multiplying by field elements first). We note that even for the case of no multiplicities, the two proofs are completely different and rely on totally different tools (ours uses additive combinatorics and [3] uses tools from real analysis). Indeed, an inherent difference between the two problems is that [3] proved that the dimension of 2-query LCCs over the reals is at most some constant whereas over finite fields the dimension can be as large as $\log_p n$ (which is, by our results, close to being best possible).

2.1. LCCs with no repetitions

Let us assume then that we have a $(2, \delta)$-LCC $V = (v_1, \ldots, v_n)$ so that no v_i and v_j are scalar multiples of each other for $i \neq j \in [n]$. We can thus treat V as a set of vectors (rather than a list). The proof has two conceptual steps. In the first step, we prove the existence of a not too small subset $V' \subseteq V$ that has low dimension. In the second step, we (iteratively) "amplify" V' until we obtain that V has low dimension.

Obtaining a (not too small) subset of low dimension: Consider the following graph on the vertex set V. We connect $v_i \sim v_j$ if there is some k such that $v_k \in \operatorname{span}(v_i, v_j)$. It is not hard to see that, by the LCC property, for every $v_k \in V$, there exists a matching M_k containing $\delta n/2$ edges, such that for every $(i,j) \in M_k$, it holds that $v_k \in \operatorname{span}(v_i, v_j)$. Assume for simplicity that it is always the case that $v_k + v_i + v_j = 0$ (we can reduce to this case by replacing each coordinate with its $p-1$ nonzero scalar multiples, we later expand on this point when discussing *LCCs in normal form*). Consider the union of all edges from all those matchings. Clearly we have $\Omega(n^2)$ edges. Label an edge (i,j) by v_k if $(i,j) \in M_k$. Notice that we have defined a *dense* graph on the vertex set V such that if $v_i \sim v_j$ then $v_i + v_j \in -V$. Intuitively, this means that the set V is "almost" a subspace. At this point, we invoke a result of Balog, Szemerédi and Gowers [2], [10] which shows that there is a not too small subset $\tilde{V} \subseteq V$ such that the size of $\tilde{V} + \tilde{V} = \{v_i + v_j : v_i, v_j \in \tilde{V}\}$ is linear in $|\tilde{V}|$, and then a result of Ruzsa [13] which implies that for such sets $\tilde{V}$, there is a not too small subset $V' \subseteq \tilde{V}$ satisfying $\dim(\operatorname{span}(V')) \leq O_{\delta,p}(1) + \log_p(n)$. Thus, in any "approximate" vector space V, a constant fraction of V spans a vector space that has almost the same size as V.

Amplification – Obtaining a (relatively large) subset of low dimension: Now we have a subset $V' \subset V$ such that $|V'|/|V| = \operatorname{poly}(\delta, p)$ and $\dim(\operatorname{span}(V')) \leq O_\delta(1) + \log_p(n)$. We would like to use induction on $V \setminus V'$ and conclude that the dimension of V is small. However, it may be the case that $|V|/|V'| > p$. In this case, the simplest argument will just give $\dim(V) < p\dim(V') = O(1) + p\log_p(n)$ which is too high (we would like the coefficient in front of the $\log_p(n)$ to only depend on δ). For that reason, we first show that we can amplify the size of V' to roughly $\delta|V|$ while increasing its dimension by only $O_{\delta,p}(1)$. The idea is that if we consider all edges labeled by elements of V', then, since there are at least $\frac{\delta}{2}|V'|n$ such edges, if $|V'| < \delta n/2$ then the induced graph on V' can only contain $|V'|^2/2 < \delta|V'|n/4$ of them. Therefore, some vertex $v \in V \setminus V'$ is adjacent to $\Omega(n)$ such edges. In particular, if we consider $V'' = V' \cup \{v\}$ and take its span, then the dimension can grow by only 1, but now, all vertices connected to v by edges whose labels come from V', also belong to V''. Thus, $|V''| \geq |V'| + \Omega(n)$. This process can continue for $O_{\delta,p}(1)$ steps and at the end we must have a set $\tilde{V}$ of size at least $\delta n/2$ and dimension $O_{\delta,p}(1) + \log_p(n)$.

Completing the argument: At this point we can consider $V \setminus \tilde{V}$ and use induction. Note that in order to use induction we must show that $V \setminus \tilde{V}$ is also a $(2, \delta')$-LCC, where $\delta' \approx \delta$. Indeed, if this is not the case then it is not hard to show that we can further increase $\tilde{V}$ by $\Omega_\delta(n)$ vertices and only increase its dimension by 1.

Concluding, since $|\tilde{V}| \geq \delta n$, we can repeat the induction at most $1/\delta$ times and get that V is the union of at most $1/\delta$ sets each of dimension at most $O_{\delta,p}(1) + \log_p(n)$. This clearly implies the result.

LCC in Normal Form: Recall that in the first step of the argument we said that without loss of generality, we assume that whenever v_i and v_j are used to recover v_k then $v_k + v_i + v_j = 0$. This is generally not the case, so what we do is, given the LCC V, we create a new LCC V' that contains all nonzero multiples (in $\mathbb{F}_p$) of every $v \in V$. In this way, whenever v_i and v_j span v_k, we can pick the appropriate multiples av_i and bv_j and get that their sum equals $-v_k$. This process, however, blows up the size of V by a factor of p, which is not too bad, but it also reduces δ to δ/p, which is a greater loss than we can afford. We therefore show in the amplification step that we can project the set that we found (which is a subset of V') back to V and get a set of density $\Omega_{\delta,p}(1)$, in V, with the required dimension.

2.2. LCC with repetitions

The argument for the case of repetitions follows the same lines, albeit the definition of a *normal form* LCC is more elaborate and the proof that a normal-form sub-code exists is considerably more involved.

Normal Form.: Given a LCC V, associate with any $v \in V$ the number $m(v)$ representing its multiplicity in V. The first step of the argument shows that given a $(2,\delta)$-LCC V, we can generate another $(2,\delta')$-LCC V' of size $n' = |V'| = \Omega_{\delta,p}(n)$ such that:

1) $\delta' = \text{poly}(\delta/p)$.
2) For every $v \in V'$, there exist $\delta' n'/2$ disjoint pairs $\{v_i, v_j\}$ such that v can be recovered from each of the pairs.
3) If v_k can be recovered from v_i and v_j, then $v_i + v_j + v_k = 0$.
4) For any two $v_i, v_j \in V'$, $m(v_i) = m(v_j)$.

We say that such V' is in normal form. In fact, what we actually do is (roughly) prove that V contains a large subset that is a LCC in normal form. This is done in Lemma 4.3, which is the main technical difficulty of the proof. Indeed the lemma shows how to reduce the case of LCCs with multiplicities to the no-multiplicity case.

Obtaining a (not too small) sublist of low dimension: We now focus on V', the LCC in normal form, that we obtained in the previous step. If we group multiples of the same vector in V' into clusters, then all the clusters are of the same size. This means that we can extract a set A of *distinct* elements, one vector from each cluster, such that A itself is an LCC. Now, we apply the Balog-Szemerédi-Gowers lemma and the Ruzsa theorem, as described in Section 2.1, to obtain a relatively large subset A' of dimension $\log_p n + O_{p,\delta}(1)$. Finally, we lift A' into a sublist V'' of V' by putting back in all the copies of vectors in A'. The lifting obviously does not change the dimension, and also because each vector has the same multiplicity, the density of A' in A and the density of V'' in V' are the same. This step is formally done in the Lemma 4.4, whose proof is in Section 5.

Amplification: Obtaining a (relatively large) sublist of low dimension.: This step is similar to the amplification step in the case of no repetitions, although it requires a slightly more careful analysis. This is given in Lemmas 4.5 and 4.6, proved in Sections 6 and 7, respectively. The end of the argument is similar to the no multiplicity case.

3. Preliminaries

3.1. Notation

Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}_p^d)^n$ be a list of n not necessarily distinct elements in $\mathbb{F}_p^d$. For a subset $S \subseteq [n]$, we denote by $V_S \in (\mathbb{F}_p^d)^{|S|}$ the sub-list of V containing all v_i's with $i \in S$. For a set $S \subseteq [n]$, we let $\text{span}_V(S) \subseteq [n]$ be defined as

$$\text{span}_V(S) = \{i \in [n] \mid v_i \in \text{span}(V_S)\}.$$

If $S = \{i\}$ is a singleton set, then we let $\text{span}_V(i) = \text{span}_V(\{i\})$. We refer to a subset $M \subseteq A \times A$ of some product set as a *matching* if for every $(i,j) \neq (i',j') \in M$ it holds that $|\{i, i', j, j'\}| = 4$. For two vectors $v, u \in \mathbb{F}_p^d$, we denote by $\text{span}(v,u) = \{av + bu \mid a, b \in \mathbb{F}_p\}$ and $\text{span}^*(v,u) = \{av + bu \mid a, b \in \mathbb{F}_p^*\}$. We will often use the simple fact that if $w \in \text{span}^*(v,u)$, then $u \in \text{span}^*(v,w)$. For a list of elements $\ell = (a_1, \ldots, a_n) \in A^n$ and an element $b \in A$, we denote by $m_\ell(b)$ the number of times b appears in ℓ (i.e., the *multiplicity* of b in ℓ).

3.2. Additive Combinatorics

For a set A in a commutative group we denote $A - A = \{a_1 - a_2 \mid a_1, a_2 \in A\}$. We will need a slight generalization of a result known as the Balog-Szemerédi-Gowers Lemma.

Theorem 3.1 ([2], [10]). *Let $\epsilon > 0$ and let $A, B \subseteq \mathbb{F}_p^d$. Suppose that there are $\epsilon|A|^2$ pairs of elements $(a,b) \in A^2$ such that $a + b \in B$. Then there exists a subset $A' \subseteq A$ with $|A'| \geq (\epsilon/2)|A|$ and such that $|A' - A'| \leq (4/\epsilon)^8 |B|^4/|A|^3$.*

Since the above statement is slightly different from the one appearing in the literature, we reprove Theorem 3.1 in the full version of this paper [6].

Another result from additive combinatorics that we will use is the following theorem of Ruzsa.

Theorem 3.2 ([13]). *Let $A \subseteq \mathbb{F}_p^d$ be such that $|A - A| \leq K|A|$. Then, there exists a subspace W of $\mathbb{Z}_p^d$ containing A such that $|W| \leq K^2 \cdot p^{K^4}|A|$. In particular, we get that*

$$\dim(W) = \log_p |W| \leq 2K^4 + \log_p |A|.$$

3.3. A useful lemma

The following simple lemma will be used several times in the proofs to follow.

Lemma 3.3. *Let $V = (v_1, \ldots, v_n) \subseteq (\mathbb{F}_p^d)^n$ be a $(2, \delta)$-LCC such that for all $v_i \in V$, $|\text{span}_V(i)| < \gamma n$. Then there exist n matchings $M_1, \ldots, M_n \subseteq [n]^2$, with $|M_k| \geq (\delta - 2\gamma)n/2$ for all $k \in [n]$, such that for every $k \in [n]$ and for every edge $(i, j) \in M_k$, $v_k \in \text{span}^*(v_i, v_j)$ and $v_k \notin \text{span}^*(v_i) \cup \text{span}^*(v_j)$.*

Proof: To see why these matchings exist, consider the following simple process of constructing them: For each $k \in [n]$, add to M'_k an edge (i, j) such that $v_k \in \text{span}(v_i, v_j)$. By the LCC property, as long as $|M'_k| \leq (\delta/2)n$, there will be another edge that we can add that does not touch any of the edges that we already added. Note that at most γn of the pairs in M'_k can contain a multiple of v_k as an element. Let $M_k \subseteq M'_k$ consist of all pairs not involving a constant multiple of v_k. It is clear that M_k has the required properties. ■

4. Proof of Theorem 1

In this section, we give the proof of Theorem 1. We first state some lemmas that will be essential for the proof. For the sake of readability, we postpone the proofs of most lemmas to later sections. For the rest of this section, let $V = (v_1, \ldots, v_n) \in (\mathbb{F}_p^d)^n$ denote a $(2, \delta)$-LCC and $\epsilon > 0$ be a sufficiently small constant.

The heart of the proof of Theorem 1, as described in Section 2.2, is the next lemma that guarantees that we can find a subset of V which is not too small and that has a low dimension.

Lemma 4.1 (Small Subset Lemma)**.** *There exist constants $c_3, c_4 > 0$ such that the following holds. Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}_p^d)^n$ be a $(2, \delta)$-LCC. Then there exists $S \subseteq [n]$ with $|S| \geq \mu(\delta, p) \cdot n$ such that*

$$\dim(V_S) \leq 1/\mu(\delta, p) + \log_p(n),$$

where $\mu(\delta, p) = \left(c_3(p/\delta)^{c_4}\right)^{-1}$.

Proof: The proof is composed of two parts. First, we show that in any LCC, we can find a smaller code that has a "nicer" structure that we call a *normal form*.

Definition 4.2 (Normal-form LCC)**.** *Let $U = (u_1, \ldots, u_n) \in (\mathbb{F}_p^d)^n$. We say that U is a* normal-form *$(2, \delta)$-LCC if there is a simple graph G with vertex set $[n]$ and with each edge labeled by some integer in $[n]$ such that the following conditions hold.*

1) *For each $k \in [n]$, the edges labeled k contain a matching consisting of δn edges.*
2) *For an edge (i, j) with label k, it holds that $u_i + u_j + u_k = 0$.*
3) *For every pair of vertices $i, j \in [n]$, we have $m_U(u_i) = m_U(u_j)$. In other words, all vertices in U have the same multiplicities.*

It might not be very obvious from the definition, but one of the main advantages of a normal from LCC stems from the fact that the graph G is **simple**. This corresponds to saying that each pair of coordinates is used in the decoding of only a **single** coordinate of the LCC. This property is easy to ensure if there are no repetitions, but is very hard to obtain otherwise, since many copies of the same vector might all 'want' to use the same edge to decode themselves, and we must decide what copy will use what edge.

The following argument shows that if no vector appears with too high a multiplicity, then we can find a subcode which is in normal form. Assume without loss of generality that for any $i, j \in [n]$, if v_i and v_j are linearly dependent, then in fact $v_i = v_j$. (Indeed this is easy to achieve by rescaling each vector, if necessary) Now, we "blow up" the code to contain all constant multiples of each coordinate. For each $v_i \in V$, let

$$L(v_i) = (v_i, 2v_i, \ldots, (p-1)v_i)$$

be the list of length $p - 1$ containing all constant multiples of v_i (except the zero one). Let V' denote the concatenation of all the lists $L(v_i)$, where $i \in [n]$. In particular, V' is a list, of size $n' = |V'| = n(p-1)$, of vectors in $\mathbb{F}_p^d$, and for any $i \in [n]$ and $c \in \mathbb{F}_p^*$, $m_V(v_i) = m_{V'}(cv_i)$. Let us denote $V' = (v'_1, \ldots, v'_{n'})$. The next lemma, shows that V' contains a sub-list which is an LCC in normal form. This is the main technical step of the proof. Due to its length, the proof of this lemma is omitted and can be found in the full version of the paper [6].

Lemma 4.3 (Subcode in Normal Form)**.** *Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}_p^d)^n$ be a $(2, \delta)$-LCC, and let V' be defined as above. If no vector $v \in V$ satisfies $m_V(v) \geq \delta^2 n/16$, then there exists a set $T \subseteq [n']$ with $|T| = t \geq \alpha \cdot n'$ such that V'_T is a normal-form $(2, \alpha)$-LCC, where $\alpha = (\delta/100p)^6$.*

The next lemma shows that if V is in normal form, then we can find a not too small subcode in it that has low dimension.

Lemma 4.4 (Small Subset Lemma for Normal Form Codes)**.** *There exist constants $c_5, c_6 > 0$ such that the following holds. Let $U = (u_1, \ldots, u_t) \in (\mathbb{F}_p^d)^t$ be a $(2, \alpha)$-LCC in normal form. Then there exists a set $S \subseteq [t]$ with $|S| \geq \tilde{\mu}(\alpha, p) \cdot t$ such that*

$$\dim(U_S) \leq 1/\tilde{\mu}(\alpha, p) + \log_p(t),$$

with $\tilde{\mu}(\alpha, p) = \left(c_5(p/\alpha)^{c_6}\right)^{-1}$.

We defer the proof of Lemma 4.4 to Section 5 and continue with the proof of Lemma 4.1.

Consider two cases. If there is $v_i \in V$ such that $m_V(v_i) \geq \delta^2 n/16$, then we define $S = \mathrm{span}_V(i)$. Clearly, $|S| \geq \delta^2 n/16$ and $\dim(S) = 1$. Thus, S is the required set. On the other hand, if for all $v_i \in V$, $m_V(v_i) < \delta^2 n/16$, then Lemma 4.3 guarantees that there is $T \subseteq [n']$ with $|T| = t \geq \alpha \cdot n' = \alpha(p-1)n$, such that V'_T is a normal-form $(2, \alpha)$-LCC, where $\alpha = (\delta/100p)^6$. By Lemma 4.4 we get that there exists a set $S' \subseteq [t]$ with $|S'| \geq \tilde{\mu}(\alpha, p) \cdot t \geq \tilde{\mu}(\alpha, p)\alpha(p-1)n$ of dimension

$$\dim(V_{S'}) \leq 1/\tilde{\mu}(\alpha, p) + \log_p(t) \leq 1/\mu(\delta, p) + \log_p(n),$$

where $\mu(\delta, p) = (c_3(p/\delta)^{c_4})^{-1}$, for some constants $c_3, c_4 > 0$. We now let $S \subset [n]$ be the set of indices of all vectors v_i that are a constant multiple of an element (whose index is) in S'. It follows that S has the required properties since its size can drop by a factor of p and its dimension stays the same. ■

Our next step is obtaining a subset of V of size roughly δn that has dimension $O_{p,\delta}(1) + \log_p(n)$. This "amplification" is guaranteed by the next lemma, whose proof applies Lemma 4.1 iteratively.

Lemma 4.5 (Large Subset Lemma). *Let $\epsilon > 0$ be a small enough constant. There exist constants $c_7, c_8 > 0$ such that the following holds. Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}_p^d)^n$ be a $(2, \delta)$-LCC. Then, there exists a set $S \subseteq [n]$ with $|S| \geq (\delta - \epsilon\delta^{1.5})n$ such that*

$$\dim(V_S) \leq \eta(\epsilon, \delta, p) + \log_p(n),$$

where $\eta(\epsilon, \delta, p) = (\epsilon\delta^3\mu(\delta/3, p)/33)^{-1} = c_7(p/\epsilon\delta)^{c_8}$.

The final lemma that we state before giving the proof of Theorem 1 shows that once we have found a subset $S \subseteq [n]$ such that $\mathrm{span}_V(S) = S$, then we can add to S some $\Omega(\delta n)$ new (indices of) vectors from V while increasing its dimension by only $O(1) + \log_p(n)$. In this fashion, we will be able to "grow" S until it equals all of $[n]$.

Lemma 4.6. *Let $\epsilon > 0$ be a small enough constant. Suppose $S \subseteq [n]$ is such that $\mathrm{span}_V(S) = S$ and $S \neq [n]$. Then there is a set $S \subseteq S' \subseteq [n]$ with $\mathrm{span}_V(S') = S'$ such that*

1) *Either $S' = [n]$ or $|S'| \geq |S| + (\delta/(2+\epsilon))n$.*
2) *$\dim(V_{S'}) \leq \dim(V_S) + \eta(\epsilon/10, \delta/3, p) + \log_p(n)$, where $\eta(\epsilon, \delta, p)$ is defined in Lemma 4.5.*

We again postpone the proofs of both Lemmas 4.5 and 4.6 (to Sections 6 and 7, respectively) and instead give the proof of Theorem 1.

Proof of Theorem 1: Let $V = (v_1, \ldots, v_n) \in (\mathbb{F}_p^d)^n$ be a $(2, \delta)$-LCC. We now apply Lemma 4.6 iteratively. Start with $S_1 = \emptyset$ and apply Lemma 4.6 repeatedly to obtain sets $S_2, S_3, \ldots$, such that for all i,

$$|S_i| \geq |S_{i-1}| + (\delta/(2+\epsilon))n$$

and

$$\dim(S_i) \leq \dim(S_{i-1}) + \eta(\epsilon/10, \delta/3, p) + \log_p(n).$$

Since the size of S_i cannot grow beyond n, the process will terminate after at most $m = \lfloor (2+\epsilon)/\delta \rfloor$ steps, yielding $S_m = [n]$. We then get that

$$\begin{aligned} \dim(V_{S_m}) &= \dim(V) \leq ((2+\epsilon)/\delta)\eta(\epsilon/10, \delta/3, p)) \\ &+ ((2+\epsilon)/\delta) \cdot \log_p(n), \end{aligned}$$

as required. This completes the proof of Theorem 1. ■

5. Proof of Lemma 4.4

Let $U = (u_1, \ldots, u_t)$ be a $(2, \alpha)$-LCC in normal form. Let G be the labeled graph on vertex set $[t]$ satisfying the requirements of the definition of normal-form LCC (Definition 4.2). Notice that G has at least αt^2 edges since there are at least αt edges for each label in $[t]$ and each edge has a unique label. Recall also that the graph G is simple (i.e. does not have repeated edges nor self loops). Also, for any two vertices i, j in G, we have that $m_U(u_i) = m_U(u_j) = m$ (say).

We can thus partition the vertices of G into $K = t/m$ disjoint sets $C_1, \ldots, C_K$ such that each C_i contains all vertices in G with the same associated vector.

Let G' be the graph obtained from G by contracting each of the sets $C_1, \ldots, C_K$ to a single vertex and erasing parallel edges and self loops.

Claim 5.1. *G' has t/m vertices and at least $\gamma \cdot (t/m)^2$ edges, where $\gamma = \alpha/4$.*

Proof: Since G is simple, the number of edges between any two sets C_i and $C_{i'}$ (including edges inside each set) can be bounded by

$$(|C_i| + |C_{i'}|)^2 = 4m^2.$$

Therefore, the number of edges in H' can decrease by at most this factor. Since the original number of edges before the contraction was at least αt^2, the number of edges remaining is at least

$$\frac{\alpha t^2}{4m^2} = \gamma \cdot (t/m)^2.$$

The calculation of the number of vertices in G' follows from the facts that each $|C_i|$ has size m and that the total number of vertices before the contraction is at most t. ■

We would now like to use Theorem 3.1 (Balog-Szemerédi-Gowers theorem). Since the sets C_i before the contraction consisted of repetitions of the same vector in U, each vertex in G' has a *distinct* vector in $\mathbb{F}_p^d$ associated with it. Let $A \subseteq \mathbb{F}_p^d$ denote the set of distinct elements $\{-u_i \mid i \in [t]\}$ and $B \subseteq \mathbb{F}_p^d$ the set of distinct elements $\{u_i \mid i \in [t]\}$. Clearly, $|A| = |B| = t/m$ by Claim 5.1. Notice that the labeling of G induces a labeling of G' since, if two edges in G have their endpoints in the same two sets C_i and $C_{i'}$ then they necessarily have labels corresponding to (repetitions of) the same vector in U (this follows from Item 2 in Definition 4.2). Thus, each edge (i_1, i_2) of G'

labeled by i_3 produces a pair of elements $(-u_{i_1}, -u_{i_2}) \in A$ such that $(-u_{i_1})+(-u_{i_2}) = u_{i_3} \in B$. Since there are at least $\gamma(t/m)^2$ distinct edges, there are $\gamma(t/m)^2 \geq \gamma \cdot |A|^2$ many such distinct pairs in A^2. We can now apply Theorem 3.1 to find a subset $A' \subseteq A$ of size $|A'| \geq (\gamma/2)|A|$ such that

$$|A' - A'| \leq (4/\gamma)^8 |B|^4/|A|^3. \tag{1}$$

Using $|A| = |B|$ and $|A| \leq (2/\gamma)|A'|$:

$$|A' - A'| \leq (4/\gamma)^9 |A'|.$$

We now apply Ruzsa's Theorem (Theorem 3.2) and conclude that A' is contained in a subspace $W \subseteq \mathbb{F}_p^d$ of dimension at most

$$\dim(W) \leq \text{poly}(1/\gamma) + \log_p |A'| \leq \text{poly}(1/\gamma) + \log_p(t).$$

Our final step is to 'lift' the set A' into a subset $S \subseteq [t]$ that will satisfy the conditions of Lemma 4.4. Let $S \subseteq [t]$ be the subset consisting of indices of vectors in U that are equal to a vector in A'. Since in the contraction step (going from G to G'), each vector was of multiplicity m, we get that

$$|S| = |A'| \cdot m \geq (\gamma m/2) \cdot |A| = \gamma t/2.$$

It is also clear that the dimension of U_S is the same as that of A'. This completes the proof of Lemma 4.4. □

6. Proof of Lemma 4.5

Let $V = (v_1, \ldots, v_n)$ be a $(2, \delta)$-LCC as in the statement of the lemma. The proof will use Lemma 4.1 as a black box, iteratively. To facilitate the iteration process we start by proving the following claim.

Claim 6.1. *Let $\epsilon > 0$ be sufficiently small and $\delta' > (\delta - \epsilon\delta^{1.5})/2$. Let $S \subseteq [n]$ be some (possibly empty) set and denote $S^c = [n] \setminus S$. Suppose that for every $k \in S^c$ there exists a matching $M_k \subseteq S^c \times S^c$ of size $\delta' n$ such that for every $(i,j) \in M_k$, $v_k \in \text{span}^*(v_i, v_j)$. Then, there exists a set $T \subseteq S^c$ and $\delta'' > 0$ such that*

1) *$|T| \geq (\delta - \epsilon\delta^{1.5})\mu(\delta', p)n$, where $\mu(\delta, p)$ is given by Lemma 4.1.*
2) *$\dim(V_T) \leq (\epsilon\delta^3\mu(\delta', p)/33)^{-1} + \log_p(n)$.*
3) *$\delta'' \geq \delta' - (\epsilon\delta^3/32)\mu(\delta', p)$.*
4) *For every $k \in S^c \setminus T$ there exists a matching $N_k \subseteq (S^c \setminus T) \times (S^c \setminus T)$ of size $\delta'' n$ such that for every $(i,j) \in N_k$, $v_k \in \text{span}^*(v_i, v_j)$. The set $S^c \setminus T$ might be empty (in which case this condition is trivially satisfied).*

Roughly, the claim says that if after removing a set S from the LCC the remaining vectors in S^c also form a (possibly slightly weaker) LCC then we can continue and 'peel' a (relatively large) subset T of S^c that has a low dimension such that $S^c \setminus T$ is also a LCC with roughly the same parameters as S^c.

Proof of Claim 6.1: Let $U = V_{S^c}$ and denote the size of the list U by $n_1 = |S^c|$. Observe that since S^c contains matchings of size $\delta' n$ and $\delta' > (\delta - \epsilon\delta^{1.5})/2$ we get that

$$n_1 \geq 2\delta' n > (\delta - \epsilon\delta^{1.5})n. \tag{2}$$

From the condition on the matchings M_k it follows that U is a $(2, \delta')$-LCC. Lemma 4.1 implies that there exists a set $T' \subseteq S^c$ such that

$$|T'| \geq \mu(\delta', p)n_1 > (\delta - \epsilon\delta^{1.5})\mu(\delta', p)n$$

and

$$\dim(U_{T'}) = \dim(V_{T'}) \leq \mu(\delta', p)^{-1} + \log_p(n).$$

Without loss of generality, we can assume that

$$\text{span}_U(T') = T'$$

(otherwise replace T' with $\text{span}_U(T')$). We will now add a small number of elements to T' to get the set T required by the claim.

Let $R = S^c \setminus T'$. Suppose that there exists some $k \in R$ such that Condition 4 of the claim does not hold (for δ'' as in Condition 3 of the claim). This means that, in the matching M_k, there are at least

$$m \geq (\epsilon\delta^3/32)\mu(\delta', p)n$$

pairs, call them

$$(i_1, j_1), \ldots, (i_m, j_m) \in U \times U$$

such that each pair contains at least one element of T', say it is always the first coordinate. Since $k \notin \text{span}_U(T')$ we know that no pair can have both its elements in T' (if this happens then v_k is spanned by elements in $V_{T'}$) and so $j_1, \ldots, j_m$ are not in T'. Therefore, by replacing T' with $\text{span}_U(T' \cup \{k\})$ we increase the size (of T') by at least m, since we are adding all the elements $j_1, \ldots, j_m$ that were not in T' before (here we use the fact that if $v_k \in \text{span}^*(v_i, v_j)$ then $v_j \in \text{span}^*(v_i, v_k)$). This step can increase the dimension by at most one. We can repeat this process at most

$$\lfloor n/m \rfloor \leq \left\lfloor ((\epsilon\delta^3/32)\mu(\delta', p))^{-1} \right\rfloor$$

times (since the size of T' cannot exceed n) and so after we are done we have a set T that satisfies Conditions 4 and 3 of the claim. Since we only added elements to T', Condition 1 is also satisfied. Condition 2 follows from the fact that at each step we increase the dimension by one and so

$$\begin{aligned}\dim(V_T) &\leq \dim(V_{T'}) + \lfloor n/m \rfloor \\ &\leq (\epsilon\delta^3\mu(\delta', p)/32)^{-1} + \mu(\delta', p)^{-1} + \log_p(n) \\ &\leq (\epsilon\delta^3\mu(\delta', p)/33)^{-1} + \log_p(n),\end{aligned}$$

where the last inequality holds for a small enough ϵ. ■

We now continue with the proof of Lemma 4.5. As before we assume that any two vectors in V are either equal or linearly independent. Set $S_0 = \emptyset$. As long as there is $k \in [n]$ with $|\mathrm{span}_V(k)| = m_V(k) \geq \epsilon\delta^2 n/16$, add k to S_0. Clearly this process terminates after at most $16/\epsilon\delta^2$ steps resulting in a set S_0 of dimension at most $16/\epsilon\delta^2$. Assume without loss of generality that $S_0 = \mathrm{span}_V(S_0)$ (otherwise we can simply increase S_0). Clearly, each $k \in [n] \setminus S_0$ has $|\mathrm{span}_V(k)| < \epsilon\delta^2 n/16$. Using the same argument as in Lemma 3.3, we conclude that there are $n_0 \triangleq n - |S_0|$ matchings $M_1^1, \ldots, M_{n_0}^1 \subseteq [n]^2$ such that $|M_k^1| \geq (\delta - \epsilon\delta^2/8)n/2$ for all $k \in [n] \setminus S_0$, and every pair $(i,j) \in M_k^1$ is so that $v_k \in \mathrm{span}^*(v_i, v_j)$. Now, if there is $k \in [n] \setminus S_0$ such that at least $\epsilon\delta^2 n/16$ of the edges in M_k^1 involve an element of S_0, then we add k to S_0 and again, take the span of the set. As in the proof of Claim 6.1, the span will contain at least $\epsilon\delta^2 n/16$ new elements. We repeat this process until we cannot continue anymore. Since the size increases at every step by at least $\epsilon\delta^2 n/16$, whereas the dimension increases by only 1, the final set, which we denote by S_1, has dimension at most $32/\epsilon\delta^2$. If $|S_1| \geq (\delta - \epsilon\delta^{1.5})n$, then we let $S = S_1$ and we are done. So assume that $|S_1| < (\delta - \epsilon\delta^{1.5})n$. At this point, each element $k \in [n] \setminus S_1$ has multiplicity smaller than $\epsilon\delta^2 n/16$ and at least $(\delta - \epsilon\delta^2/4)n/2$ edges in M_k^1 do not involve any element of S_1.

We would like to apply Claim 6.1 with S_1 being the set S of the claim. Before doing so we set

$$\delta_1 = (\delta - \epsilon\delta^2/4)/2,$$

and note that for each $k \in S_1^c$, at least $(\delta - \epsilon\delta^2/4)n/2 = \delta_1 n$ of the edges in M_k^1 do not involve any element of S_1. We can now apply Claim 6.1 with $\delta' = \delta_1 = (\delta - \epsilon\delta^2/4)/2 > (\delta - \epsilon\delta^{1.5})/2$ to find a subset $T_1 \subseteq S_1^c$ which satisfies the conditions of the claim. In particular

$$|T_1| \geq (\delta - \epsilon\delta^{1.5})\mu(\delta_1, p)n \geq (\delta - \epsilon\delta^{1.5})\mu(\delta/4, p)n$$

and

$$\begin{aligned} \dim(V_{T_1}) &\leq (\epsilon\delta^3\mu(\delta_1, p)/33)^{-1} + \log_p(n) \\ &\leq (\epsilon\delta^3\mu(\delta/4, p)/33)^{-1} + \log_p(n). \end{aligned}$$

We also get, for every $k \in S_1^c \setminus T_1$, a new matching M_k^2 that satisfies Condition 4 of Claim 6.1 and whose size is

$$\begin{aligned} |M_k^2| &\geq \delta'' n \\ &\geq (\delta_1 - (\epsilon\delta^3/32)\mu(\delta_1, p))n \\ &\geq ((\delta - \epsilon\delta^2/4)/2 - (\epsilon\delta^3/32)\mu(\delta, p))n \\ &> (\delta - \epsilon\delta^{1.5})n/2. \end{aligned}$$

Set $\delta_2 = \delta'' > (\delta - \epsilon\delta^{1.5})/2$. Let $S_2 = S_1 \cup T_1$. We can now apply Claim 6.1 with $S = S_2$. This process will result in a sequence of disjoint sets $T_1, T_2, \ldots$ and corresponding matchings $\{M_k^1\}, \{M_k^2\}, \ldots$ of sizes $\delta_1 n, \delta_2 n, \ldots$ where $\delta_{i+1} \geq \delta_i - (\epsilon\delta^3/32)\mu(\delta_i, p) \geq \delta_i - (\epsilon\delta^3/32)\mu(\delta, p)$. We will also have the related sequence of sets

$$S_1, S_2, \ldots, S_i = S_{i-1} \cup T_{i-1}.$$

We will stop at step ℓ if we get $\delta_\ell \leq (\delta - \epsilon\delta^{1.5})/2$ or if we run out of elements of $[n]$ (that is, if $S_\ell = [n]$).

Suppose this process stops after ℓ iterations. Since we have found ℓ disjoint sets $T_1, \ldots, T_\ell$, each of size at least $(\delta - \epsilon\delta^{1.5})\mu(\delta/4, p)n$ it holds that

$$\ell \leq \left\lfloor \left((\delta - \epsilon\delta^{1.5})\mu(\delta/4, p)\right)^{-1} \right\rfloor.$$

We can use the bound on ℓ to obtain

$$\begin{aligned} \delta_\ell &\geq \delta_1 - (\ell - 1) \cdot (\epsilon\delta^3/32)\mu(\delta/4, p) \\ &> (\delta - \epsilon\delta^2/4)/2 \\ &\quad - \left((\delta - \epsilon\delta^{1.5})\mu(\delta/4, p)\right)^{-1} \cdot (\epsilon\delta^3/32)\mu(\delta/4, p) \\ &> (\delta - \epsilon\delta^{1.5})/2 \end{aligned}$$

and so the process will terminate only after we covered all of $[n]$. Notice that, as the process did not terminate at the $(\ell - 1)$'th step, it must be the case that

$$|S_{\ell-1}| \leq (1 - (\delta - \epsilon\delta^{1.5}))n$$

since, otherwise, the set $[n] \setminus S_{\ell-1}$ would not be big enough to contain the matchings $\{M_k^{\ell-1}\}$ which have at least $\delta_{\ell-1} n > (\delta - \epsilon\delta^{1.5})n/2$ edges each. This implies that

$$|T_{\ell-1}| = |S_\ell| - |S_{\ell-1}| \geq (\delta - \epsilon\delta^{1.5})n.$$

The proof of Lemma 4.5 is now complete since, by Condition 2 of Claim 6.1, we have

$$\begin{aligned} \dim(V_{T_{\ell-1}}) &\leq (\epsilon\delta^3\mu(\delta_{\ell-1}, p)/33)^{-1} \\ &+ \log_p(n) \leq (\epsilon\delta^3\mu(\delta/4, p)/33)^{-1} \\ &+ \log_p(n). \end{aligned}$$

7. Proof of Lemma 4.6

The proof of this lemma is similar to Proposition 7.11 in [3].

Let $S^c = [n] \setminus S$. As in the proof of Lemma 4.5, we first add to S all elements $k \in S^c$ with $|\mathrm{span}_V(k)| \geq \epsilon\delta^2 n/20$ and denote by S_1 the span of the resulting set. This process can add at most $20/\epsilon\delta^2$ linearly independent elements to S and so $\dim(S_1) \leq \dim(S) + 20/\epsilon\delta^2$. We again follow the argument of Lemma 3.3 and conclude that for every $k \in [n] \setminus S_1$, there is a matching $M_k \subseteq [n]^2$, of size $|M_k| \geq (\delta - \epsilon\delta^2/10)n/2$, such that for each $(i,j) \in M_k$ we have $v_k \in \mathrm{span}^*(v_i, v_j)$. We now repeat the following: We add to S_1 any k such that M_k contains at least $\epsilon\delta^2 n/20$ edges with at least one endpoint in S_1 and take the span (inside V) of this set. It is clear that whenever we add such an element to S_1 its size grows by $\epsilon\delta^2 n/20$ and its dimension grows by 1. Thus, this process ends after at most $20/\epsilon\delta^2$ steps. Call the resulting set S_2. If $S_2 = [n]$, then we set $S' = S_2$ and complete the proof. Otherwise, since $S_2 \neq [n]$, there must be $k \in S_2^c$. As M_k has $(\delta - \epsilon\delta^2/5)n/2$ edges in $S_2^c \times S_2^c$ (as otherwise we would have added v to S_2), it must be the case that $|S_2^c| \geq (\delta - \epsilon\delta^2/5)n$.

Denote $n_2 = |S_2^c|$. From the argument above, it follows that there are n_2 matchings $\{M'_k\}_{k \in S_2^c}$, with $M_k \subseteq (S_2^c)^2$, such that for all $k \in S_2^c$, $|M_k| \geq (\delta - \epsilon\delta^2/5)n/2$ and for each $(i,j) \in M_k$ we have $v_k \in \text{span}^*(v_i, v_j)$. This implies that

$$V' = V_{S^c}$$

is a $(2, \delta')$-LCC with

$$\delta' = (1/2)(\delta - \epsilon\delta^2/5)(n/n_2).$$

Indeed, we get such δ' since for every $k \in S_2^c$, $|M_k| \geq (\delta - \epsilon\delta^2/5)n/2 \geq \delta' n_2$. Lemma 4.5 now implies that there is a subset $\widehat{S} \in S^c$ such that

$$\begin{array}{rcl} |\widehat{S}| & \geq & (\delta' - \epsilon\delta'^{1.5}/10)n_2 \\ & \geq & (1 - \epsilon/10)\delta' n_2 \\ & \geq & (1 - \epsilon/3)\delta n/2 \\ & \geq & \delta n/(2 + \epsilon) \end{array}$$

and

$$\begin{array}{rcl} \dim(V_{\widehat{S}}) & \leq & \eta(\epsilon/10, \delta', p) + \log_p(n) \\ & \leq & \eta(\epsilon/10, \delta/3, p) + \log_p(n). \end{array}$$

Letting

$$S' = \text{span}_V(S \cup \widehat{S})$$

completes the proof of Lemma 4.6. □

8. Acknowledgement

A.B. research conducted as a graduate student at MIT and partially supported by NSF Awards 0514771, 0728645, and 0732334.

Z.D. research partially supported by NSF grant CCF-0832797 and by the Packard fellowship.

S.S. research partially supported by the Microsoft Research Ph.D. Fellowship.

Part of A.S. research was done while visiting MSR NE. A.S. research was partially supported by the Israel Science Foundation (grant number 339/10).

References

[1] N. Alon, "Perturbed identity matrices have high rank: Proof and applications," *Combin. Probab. Comput.*, vol. 18, no. 1-2, pp. 3–15, 2009.

[2] A. Balog and E. Szemerédi, "A statistical theorem of set addition," *Combinatorica*, vol. 14, pp. 263–268, 1994.

[3] B. Barak, Z. Dvir, A. Wigderson, and A. Yehudayoff, "Rank bounds for design matrices with applications to combinatorial geometry and locally correctable codes," in *Proc. 43rd Annual ACM Symposium on the Theory of Computing (to appear)*, 2011.

[4] O. Barkol, Y. Ishai, and E. Weinreb, "On locally decodable codes, self-correctable codes, and t-private PIR," *Algorithmica*, vol. 58, pp. 831–859, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00453-008-9272-1

[5] J. Beck, "On the lattice property of the plane and some problems of Dirac, Motzkin and Erdős in combinatorial geometry," *Combinatorica*, vol. 3, pp. 281–297, 1983.

[6] A. Bhattacharyya, Z. Dvir, S. Saraf, and A. Shpilka, "Tight lower bounds for 2-query LCCs over finite fields," 2011, ECCC Technical-Report TR11-054. [Online]. Available: http://eccc.hpi-web.de/report/2011/054/

[7] Z. Dvir, "On matrix rigidity and locally self-correctable codes," in *Proc. 25th Annual IEEE Conference on Computational Complexity*, 2010, pp. 291–298.

[8] Z. Dvir and A. Shpilka, "Locally decodable codes with two queries and polynomial identity testing for depth 3 circuits," *SIAM J. Comput.*, vol. 36, no. 5, pp. 1404–1434, 2007.

[9] O. Goldreich, H. J. Karloff, L. J. Schulman, and L. Trevisan, "Lower bounds for linear locally decodable codes and private information retrieval," *Comput. Complexity*, vol. 15, no. 3, pp. 263–296, 2006.

[10] T. Gowers, "A new proof of Szemerédi's theorem for arithmetic progressions of length four," *Geom. Funct. Anal.*, vol. 8, pp. 529–551, 1998.

[11] N. Kayal and S. Saraf, "Blackbox polynomial identity testing for depth 3 circuits," in *Proceedings of the 50th Annual FOCS*, 2009, pp. 198–207.

[12] I. Kerenidis and R. de Wolf, "Exponential lower bound for 2-query locally decodable codes via a quantum argument," *J. Comput. System Sci.*, vol. 69, no. 3, pp. 395–420, 2004.

[13] I. Ruzsa, "Sums of finite sets," in *Number Theory: New York Seminar*, D. V. Chudnovsky, G. V. Chudnovsky, and M. B. Nathanson, Eds. Springer Verlag, 1996.

[14] N. Saxena and C. Seshadhri, "From Sylvester-Gallai configurations to rank bounds: Improved black-box identity test for depth-3 circuits," in *Proc. 51st Annual IEEE Symposium on Foundations of Computer Science*, 2010, pp. 21–29.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[16] S. Yekhanin, "Locally decodable codes," *Foundations and Trends in Theoretical Computer Science*, to appear. Preliminary version at http://research.microsoft.com/en-us/um/people/yekhanin/Papers/LDCnow.pdf.