Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 149 (2010)

Rank Bounds for Design Matrices with Applications to Combinatorial Geometry and Locally Correctable Codes

Boaz Barak* Zeev Dvir† Avi Wigderson‡ Amir Yehudayoff§

Abstract

A (q, k, t)-design matrix is an $m \times n$ matrix whose pattern of zeros/non-zeros satisfies the following design-like condition: each row has at most q non-zeros, each column has at least k non-zeros and the supports of every two columns intersect in at most t rows. We prove that for $m \geq n$, the rank of any (q, k, t)-design matrix over a field of characteristic zero (or sufficiently large finite characteristic) is at least

$$n - \left(\frac{qtn}{2k}\right)^2.$$

Using this result we derive the following applications:

Impossibility results for 2-query LCCs over large fields. A 2-query locally correctable code (LCC) is an error correcting code in which every codeword coordinate can be recovered, probabilistically, by reading at most two other code positions. Such codes have numerous applications and constructions (with exponential encoding length) are known over finite fields of small characteristic. We show that infinite families of such linear 2-query LCCs *do not exist* over fields of characteristic zero or large characteristic regardless of the encoding length.

Generalization of known results in combinatorial geometry. We prove a quantitative analog of the Sylvester-Gallai theorem: Let $v_1, \ldots, v_m$ be a set of points in $\mathbb{C}^d$ such that for every $i \in [m]$ there exists at least δm values of $j \in [m]$ such that the line through v_i, v_j contains a third point in the set. We show that the dimension of $\{v_1, \ldots, v_m\}$ is at most $O(1/\delta^2)$. Our results generalize to the high dimensional case (replacing lines with planes, etc.) and to the case where the points are colored (as in the Motzkin-Rabin Theorem).

*Microsoft Research New England. Email: boaz@microsoft.com. Most of the work done while at Princeton University and supported by NSF grants CNS-0627526, CCF-0426582 and CCF-0832797, and the Packard and Sloan fellowships.

†Department of Computer Science, Princeton University. Email: zeev.dvir@gmail.com. Research partially supported by NSF grant CCF-0832797 and by the Packard fellowship.

‡School of Mathematics, Institute for Advanced Study. Email: avi@ias.edu. Research partially supported by NSF grants CCF-0832797 and DMS-0835373.

§Department of Mathematics, Technion - IIT. Email: amir.yehudayoff@gmail.com. Most of the work done while at the IAS. Research partially supported by NSF grants CCF-0832797 and DMS-0835373.

ISSN 1433-8092

1 Introduction

In this work we study what *combinatorial* properties of matrices guarantee high algebraic *rank*, where a property is combinatorial if it depends only on the zero/non-zero pattern of the matrix, and not on the values of its entries. This question has a rich history in mathematics (see Section 1.2), and some computer science motivations:

Locally correctable codes. A *locally correctable code* is an error correcting code in which for every codeword y, given a corrupted version $\tilde{y}$ of y and an index i, one can recover the correct value of y_i from $\tilde{y}$ by looking only at very few coordinates of $\tilde{y}$. It is an open question in coding theory to understand the tradeoffs between the fraction of errors, locality (number of coordinates read) and rate (ratio of message length to codeword length) of such codes, with very large gaps between the known upper bounds and lower bounds (see the survey [Tre04]). The question is open even for linear codes, where the condition of being locally correctable turns out to be equivalent to the existence of low weight codewords in the dual codewords that are "well-spread" in some precise technical sense (see Section 7). Because of the relation between the rate of the code and its dual, the question becomes equivalent to asking whether this combinatorial "well-spreadness" condition guarantees high rank.

Matrix rigidity. A longstanding question is to come up with an explicit matrix that is *rigid* in the sense that its rank cannot be reduced by changing a small number of its entries. Random matrices are extremely rigid, and sufficiently good explicit constructions will yield lower bounds for arithmetic circuits [Val77], though we are still very far from achieving this (see the survey [Lok09]). One can hope that a combinatorial property guaranteeing large rank will be robust under small perturbations, and hence a matrix satisfying such a property will automatically be rigid.

In both these cases it is crucial to obtain bounds on the rank that depend solely on the zero/non-zero pattern of the matrix, without placing any restrictions on the non-zero coefficients. For example, there are very strong bounds known for matrix rigidity under the restriction that the non-zero coefficients have bounded magnitude (see Chapter 3 in [Lok09]), but they only imply lower bounds in a very restricted model. In fact, there is a relation between the two questions, and sufficiently good answers for the first question will imply answers for the second one [Dvi10]. We stress that these two examples are in no way exhaustive. The interplay between combinatorial and algebraic properties of matrices is a fascinating question with many potential applications that is still very poorly understood.

1.1 Our Results

In this work we give a combinatorial property of complex matrices that implies high rank. While not strong enough to prove rigidity results, we are able to use it to obtain several applications in combinatorial geometry and locally correctable codes. Our main result is the following theorem, giving a lower bound on the rank of matrix whose non-zero pattern forms has certain

combinatorial-design like properties in the sense that the sets of non-zero entries in each column have small intersections. (This theorem is restated as Theorem 3.2.)

Theorem 1 (Rank bound for design matrices)**.** *Let $m \geq n$. We say that an $m \times n$ complex matrix A is a (q, k, t)-design matrix if every row of A has at most q non-zero entries, every column of A has at least k non-zeroes entries, and the supports of every two columns intersect in at most t rows. For every such A,*

$$rank(A) \geq n - \left(\frac{q \cdot t \cdot n}{2k}\right)^2.$$

We also show that Theorem 1, and in fact any result connecting the zero/non-zero pattern to rank, can be made to hold over arbitrary characteristic zero fields and also over fields of sufficiently large (depending on m, n) finite characteristic.

1.1.1 Applications to Combinatorial Geometry

Our most immediate applications of Theorem 1 are to questions regarding line-point incidences. Results on line-point incidences have recently found use in the area of computational complexity in relation to pseudo-randomness [BKT04, BIW06] and de-randomization [KS09, SS10]. In this setting we have an arrangement of a finite number of points in real or complex space. Every such arrangement gives rise to a set of lines, namely, those lines that pass through at least two of the points in the arrangement. Information about these lines can be converted, in some cases, into information about the dimension of the set of points (i.e. the dimension of the space the points span). Our rank theorem can be used to derive generalizations for two well-known theorems in this area: the Sylvester-Gallai theorem and the Motzkin-Rabin theorem.

Generalizing the Sylvester-Gallai Theorem. The Sylvester-Gallai (SG for short) theorem says that if m distinct points $v_1, \ldots, v_m \in \mathbb{R}^d$ are not collinear, then there exists a line that passes through exactly two of them. In its contrapositive form the SG theorem says that if for every $i \neq j$ the line through v_i and v_j passes through a third point v_k, then $\dim\{v_1, \ldots, v_m\} \leq 1$, where $\dim\{v_1, \ldots, v_m\}$ is the dimension of the smallest affine subspace containing the points. This theorem was first conjectured by Sylvester in 1893 [Syl93], proved (in dual form) by Melchior in 1940 [Mel40], and then independently conjectured by Erdos in 1943 [Erd43] and proved by Gallai in 1944. The SG theorem has several beautiful proofs and many generalizations, see the survey [BM90]. Over the complex numbers the (tight) bound on the dimension is 2 instead of 1. The complex version was first proven by Kelly [Kel86] using a deep results from algebraic geometry, and more recently, an elementary proof was found by Elkies, Pretorius and Swanepoel [ES06] who also proved it over the quaternions with an upper of 4 on the dimension.

We say that the points $v_1, \ldots, v_m$ (in $\mathbb{R}^d$ or $\mathbb{C}^d$) form a δ-SG configuration if for every $i \in [m]$ there exists at least δm values of $j \in [m]$ such that the line through v_i, v_j contains a third point in the set. Szemeredi and Trotter [ST83] showed that, when δ is larger than some absolute constant close to 1, then the dimension of a δ-SG configuration is at most one (over the reals). We show the following generalization of their result to arbitrary $\delta > 0$ (and over the complex numbers).

Theorem 2 (Quantitative SG theorem)**.** *If $v_1, \ldots, v_m \in \mathbb{C}^d$ is a δ-SG configuration then* $\dim\{v_1, \ldots, v_m\} < 13/\delta^2$.

We note that one cannot replace the bound $13/\delta^2$ of Theorem 2 with 1 or even with any fixed constant, as one can easily create a δ-SG configuration of dimension roughly $2/\delta$ by placing the points on $1/\delta$ lines. This is analogous to error correcting codes, where once the fraction δ of agreement between the original and corrupted codeword drops below half there can be no unique decoding. In that sense our result can be thought of as a *list decoding* variant of the SG theorem, whereas the result of [ST83] is its unique decoding variant. We also show an "average case" version of the SG theorem, proving a bound on the dimension of a large subset of the points under the assumption that there are many collinear triples (see Theorem 4.8).

We also prove a version of Theorem 4.3 with lines replaced by k-flats (k-dimensional affine subspaces). This generalizes a theorem of Hansen [Han65, BE67] which deals with the case $\alpha = 1$. The statement of this result is technical and so we give it in Section 5 where it is also proven.

Since our proofs use elementary (and purely algebraic) reductions to the rank theorem, they hold over arbitrary fields of characteristic zero or of sufficiently large finite characteristic. This is in contrast to many of the known proofs of such theorems which often rely on specific properties of the real (or complex) numbers. However, we currently do not recover the full version of the original SG theorem, in the sense that even for $\delta = 1$ we do not get a bound of 1 (or 2 for complex numbers) on the dimension. (However, the term $13/\delta^2$ can be improved a bit in the $\delta = 1$ case to obtain a bound of 9 on the dimension.)

Generalizing the Motzkin-Rabin Theorem. The Motzkin-Rabin (MR for short) theorem (see e.g. [BM90]) is an interesting variant of the Sylvester-Gallai theorem that states that if points $v_1, \ldots, v_m \in \mathbb{R}^d$ are colored either red or blue and there is no monochromatic line passing through at least two points, then they are all collinear. As in the SG theorem, we obtain a quantitative generalization of the MR theorem such that (letting b and r be the numbers of blue and red points respectively), if for every blue (resp. red) point v, there are δb blue (resp. δr red) points v' where the line through v and v' passes through a red (resp. blue) point, then $\dim\{v_1, \ldots, v_m\} \leq O(1/\delta^4)$. We also prove a three colors variant of the MR theorem, showing that if $v_1, \ldots, v_m$ are colored red, blue and green, and all lines are not monochromatic, then $\dim\{v_1, \ldots, v_m\}$ is at most some absolute constant.

1.1.2 Locally Correctable Codes

A (linear) q query locally correctable code ((q, δ)-LCC for short) over a field $\mathbb{F}$ is a subspace $C \subseteq \mathbb{F}^n$ such that, given an element $\tilde{y}$ that disagrees with some $y \in C$ in at most δn positions and an index $i \in [n]$, one can recover y_i with, say, probability 0.9, by reading at most q coordinates of $\tilde{y}$. Over the field of two elements $\mathbb{F}_2$ the standard Hadamard code construction yields a (2,δ)-query LCC with dimension $\Omega(\log(n))$ for constant $\delta > 0$ (see the survey [Tre04]). In contrast we show that for every constant $\delta > 0$ there do not exist infinite family of such codes over the complex numbers:

Theorem 3 (Impossibility of 2-query LCCs over $\mathbb{C}$)**.** *If C is a 2-query LCC for δ fraction of errors over $\mathbb{C}$, then* $\dim(C) \leq O(1/\delta^9)$.

We note that the Hadamard construction does yield a *locally decodable code* over the complex numbers with dimension $\Omega(\log n)$. Locally decodable codes are the relaxation of a locally correctable codes where one only needs to be able to recover the coordinates of the original message as opposed to the codeword. Thus over the complex numbers, there is a very strong separation between the notions of locally decodable and locally correctable codes, whereas it is consistent with our knowledge that for, say, $\mathbb{F}_2$ the rate/locality tradeoffs of both notions are the same.

1.2 Related Work

The idea to use matrix scaling to study structural properties of matrices was already present in [CPR00]. This work, which was also motivated by the problem of matrix rigidity, studies the presence of short cycles in the graphs of non-zero entries of a square matrix.

A related line of work on the rank of 'design' matrices is the work emerging from Hamada's conjecture [Ham73]. (See [JT09] for a recent result and more references.) Here, a design matrix is defined using stricter conditions (each row/column has exactly the same number of non-zeros and the intersections are also all of the same size) which are more common in the literature dealing with combinatorial designs. In order to be completely consistent with this line of work we should have called our matrices 'approximate-design' matrices. We chose to use the (already overused) word 'design' to make the presentation more readable. We also note that considering approximate designs only makes our results stronger. Hamada's conjecture states that of all zero/one matrices whose support comes from a design (in the stricter sense), the minimal rank is obtained by matrices coming from geometric designs (in our language, Reed-Muller codes). In contrast to this paper, the emphasis in this line of works is typically on small finite fields. We note here that the connection between Hamada's conjecture and LCCs was already observed by Barkol, Ishai and Weinreb [BIW07] who also conjectured (over small fields) the 'approximate-design' versions which we prove here for large fields.

Another place where the support of a matrix is connected to its rank is in graph theory where we are interested in minimizing the rank of a (square, symmetric) real matrix which has the same support as the adjacency matrix of a given graph. This line of work goes back for over fifty years and has many applications in graph theory. See [FH07] for a recent survey on this topic.

Over the reals we can also ask about the minimal rank of matrices with certain *sign-pattern*. That is, given a matrix over $\{1, -1\}$, what is the minimal rank of a matrix which has the same sign-pattern. This minimal rank is called the *sign-rank* of a matrix. The question of coming up with (combinatorial or otherwise) properties that imply high sign-rank is one of major importance and has strong connections to communication complexity, learning theory and circuit complexity, among others. For a recent work with plenty of references see [RS08]. In particular we would like to mention a connection to the work of Forster [For02] on the sign-rank of the Hadamard matrix. (An earlier version of this work used a variant [Bar98, Har10] of a lemma from [For02] instead of the results of [RS89] on matrix scaling to obtain our main result.)

1.3 Organization

In Section 2 we give a high level overview of our techniques. In Section 3 we prove our main result on the rank of design matrices. In Section 4 we prove our quantitative variants of the Sylvester-Gallai theorem. In Section 5 we prove the high-dimensional analog of Theorem 4.3 where lines are replaced with flats. In Section 6 we prove our generalizations of the Motzkin-Rabin theorem. In Section 7 we prove our results on locally correctable codes. In Section 8 we show how our results extend to other fields. We conclude in Section 9 with a discussion of open problems.

2 Our Techniques

We now give high-level proof overviews for some of our results.

2.1 Rank Lower Bounds for Design Matrices

Theorem 1 – the rank lower bound for design matrices – is proved in two steps. We now sketch the proof, ignoring some subtleties and optimizations. The proof starts with the observation that, as in the case of matrix rigidity and similar questions, the result is much easier to prove given a bound on the *magnitude* of the non-zero entries. Indeed, if A is a (q, k, t)-design matrix and all of its non-zero entries have absolute value in $[1/c, 1]$ for some constant c, then the $n \times n$ matrix $M = A^*A$ is *diagonally dominant*, in the sense that for all $i \neq j$, $m_{ii} \geq k/c^2$ but $|m_{ij}| \leq t$. (Here A^* denotes the conjugate transpose of A.) Thus one can use known results on such matrices (e.g. [Alo09]) to argue that $\mathrm{rank}(A) \geq \mathrm{rank}(M) \geq n - (ntc^2/k)^2$. Our main idea is to reduce to this case where the non-zero coefficients of A are (roughly) bounded using *matrix scaling*.

A *scaling* $\hat{A}$ of a matrix A is obtained by multiplying for all i, j, the i'th row of A by some positive number ρ_i and the j'th column of A by some positive number γ_j. Clearly, A and $\hat{A}$ share the same rank and zero/non-zero pattern. We use known matrix-scaling results [Sin64, RS89] to show that every (q, k, t)-design matrix A has a scaling in which every entry has magnitude at most (roughly) 1 but its columns have norm at least (roughly) $\sqrt{k/q}$. We note that the typical application of matrix-scaling was with respect to the ℓ_1-norm of the rows and columns. Here we take a different path: We use scaling with respect to ℓ_2-norm.

We defer the description of this step to Section 3 but the high level idea is to use a theorem of [RS89] that shows that such a scaling exists (in fact without the dependence on q) if A had the property of not containing any large all-zero sub-matrix. While this property cannot be in general guaranteed, we show that by repeating some rows of A one can obtain a matrix B that has this property, and a scaling of B can be converted into a scaling of A. Since our lower bound on the entry m_{ii} in the bounded coefficient case (where again $M = A^*A$) only used the fact that the columns have large norms, we can use the same argument as above to lower bound the rank of M, and hence of A.

2.2 Generalized Sylvester-Gallai Theorem

Recall that the quantitative SG theorem (Theorem 2) states that every δ-SG configuration $v_1, \ldots, v_n$, has dimension at most $13/\delta^2$. Our proof of Theorem 2 uses Theorem 1 as follows. Suppose for starters that every one of these lines passed through *exactly* three points. Each such line induces an equation of the form $\alpha v_i + \beta v_j + \gamma v_k = 0$. Now for $m = \delta n^2$, let A be the $m \times n$ matrix whose rows correspond to these equations. Since every two points participate in only one line, A will be a $(3, \delta n, 1)$ design matrix, meaning that according to Theorem 1, A's rank is at least $n - \left(\frac{3}{2\delta}\right)^2$. Since A times the matrix whose rows are $v_1, \ldots, v_n$ is zero we have $\dim\{v_1, \ldots, v_n\} \leq n - \mathrm{rank}(A)$. We thus get an upper bound of $\lfloor 9/4 \rfloor = 2$ on this dimension. To handle the case when some lines contain more than three points, we choose in some careful way from each line ℓ containing r points a subset of the $\binom{r}{3}$ equations of the form above that it induces on its points. We show that at some small loss in the parameters we can still ensure the set of equations forms a design, hence again deriving a lower bound on its rank via Theorem 1.

Our method extend also to an "average case" SG theorem (Theorem 4.8), where one only requires that the set of points supports many (i.e., $\Omega(n^2)$) collinear triples and that each pair of points appear together in a few collinear triples. In this case we are able to show that there is a subset of $\Omega(n)$ points whose span has dimension $O(1)$. See Section 4 for more details. Our generalizations of the Motzkin-Rabin theorem follow from our theorem on δ-SG configurations via simple reductions (see Section 6).

2.3 Locally Correctable Codes

At first sight, Theorem 3 – non existence of 2 query locally correctable codes over $\mathbb{C}$ – seems like it should be an immediate corollary of Theorem 2. Suppose that a code C maps $\mathbb{C}^d$ to $\mathbb{C}^n$, and let $v_1, \ldots, v_n$ denote the rows of its generating matrix. That is, the code maps a message $x \in \mathbb{C}^d$ to the vector $(\langle v_1, x\rangle, \ldots, \langle v_n, x\rangle)$. The fact that C is a 2 query LCC for δ errors implies that for every such row v_i, there are roughly δn pairs j, k such that v_i is in the span of $\{v_j, v_k\}$. Using some simple scaling/change of basis, this gives precisely the condition of being a δ-SG configuration, save for one caveat: In a code there is no guarantee that all the vectors $v_1, \ldots, v_n$ are distinct. That is, the code may have repeated coordinates that are always identical. Intuitively it seems that such repetitions should not help at all in constructing LCCs but proving this turned out to be elusive. In fact, our proof of Theorem 3 is rather more complicated than the proof Theorem 2, involving repeated applications of Theorem 1 which result also in somewhat poorer quantitative bounds. The idea behind the proof to use a variant of the "average case" SG theorem to repeatedly find $\Omega(n)$ points among $v_1, \ldots, v_n$ whose span has $O(1)$ dimension, until there are no more points left. We defer all details to Section 7.

Given Theorem 1, one may have expected that Theorem 3 could be extended for LCCs of any constant number q of queries. After all, the condition of C being an LCC intuitively seems like only a slight relaxation of requiring that the dual code of C has a generating matrix whose non-zero pattern is a combinatorial design, and indeed in known constructions of LCCs, the dual code does form a design. We are not, however, able to extend our results to 3 and more queries. A partial explanation to our inability is that 3 query LCCs give rise to configuration of planes (instead of lines) and point and planes exhibit much more complicated combinatorial

properties than lines.

3 Rank of Design Matrices

In this section we prove our main result which gives a lower bound on the rank of matrices whose zero/non-zero pattern satisfies certain properties. We start by defining these properties formally.

Definition 3.1 (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 $i \in [m]$, $|R_i| \leq q$.
2. For all $j \in [n]$, $|C_j| \geq k$.
3. For all $j_1 \neq j_2 \in [n]$, $|C_{j_1} \cap C_{j_2}| \leq t$.

Theorem 3.2 (Restatement of Theorem 1 – rank of design matrices)**.** *Let A be an $m \times n$ complex matrix. If A is a (q, k, t)-design matrix then*

$$\mathit{rank}(A) \geq n - \left(\frac{q \cdot t \cdot n}{2k}\right)^2.$$

Remark 3.3. The proof of the theorem actually holds under a slightly weaker condition on the sizes of the intersections. Instead of requiring that $|C_{j_1} \cap C_{j_2}| \leq t$ for all pairs of columns $j_1 \neq j_2$, it is enough to ask that

$$\sum_{j_1 \neq j_2} |C_{j_1} \cap C_{j_2}|^2 \leq n^2 \cdot t^2.$$

That is, there could be some pairs with large intersection as long as the average of the squares is not too large.

The proof of the theorem is given below, following some preliminaries.

3.1 Preliminaries for the Proof of Theorem 3.2

Notation: For a set of real vectors $V \in \mathbb{C}^n$ we denote by $\mathrm{rank}(V)$ the dimension of the vector space spanned by elements of V. We denote the ℓ_2-norm of a vector v by $\|v\|$. We denote by I_n the $n \times n$ identity matrix.

We start with definitions and results on matrix scaling.

Definition 3.4. [Matrix scaling] Let A be an $m \times n$ complex matrix. Let $\rho \in \mathbb{C}^m, \gamma \in \mathbb{C}^n$ be two complex vectors with all entries non-zero. We denote by

$$\mathbf{SC}(A, \rho, \gamma)$$

the matrix obtained from A by multiplying the (i,j)'th element of A by $\rho_i \cdot \gamma_j$. We say that two matrices A, B of the same dimensions are a scaling of each other if there exist non-zero vectors ρ, γ such that $B = \mathbf{SC}(A, \rho, \gamma)$. It is easy to check that this is an equivalence relation. We refer to the elements of the vector ρ as the *row scaling coefficients* and to the elements of γ as the *column scaling coefficients*. Notice that two matrices which are a scaling of each other have the same rank and the same pattern of zero and non-zero entries.

Matrix scaling originated in a paper of Sinkhorn [Sin64] and has been widely studied since (see [LSW00] for more background). The following is a special case of a theorem from [RS89] that gives sufficient conditions for finding a scaling of a matrix which has certain row and column sums.

Definition 3.5 (Property-S)**.** Let A be an $m \times n$ matrix over some field. We say that A satisfies *Property-S* if for every zero sub-matrix of A of size $a \times b$ it holds that

$$\frac{a}{m} + \frac{b}{n} \leq 1. \tag{1}$$

Theorem 3.6 (Matrix scaling theorem, Theorem 3 in [RS89])**.** *Let A be an $m \times n$ real matrix with non-negative entries which satisfies Property-S. Then, for every $\epsilon > 0$, there exists a scaling A' of A such that the sum of each row of A' is at most $1 + \epsilon$ and the sum of each column of A' is at least $m/n - \epsilon$. Moreover, the scaling coefficients used to obtain A' are all positive real numbers.*

The proof of the theorem is algorithmic [Sin64]: Start by normalizing A's rows to have sum 1, then normalize A's columns to have sum m/n, then go back to normalizing the rows the have sum 1, and so forth. It can be shown (using a suitable potential function) that this process eventually transforms A to the claimed form (since A has Property-S).

We will use the following easy corollary of the above theorem.

Corollary 3.7 (ℓ_2^2-scaling)**.** *Let $A = (a_{ij})$ be an $m \times n$ complex matrix which satisfies Property-S. Then, for every $\epsilon > 0$, there exists a scaling A' of A such that for every $i \in [m]$*

$$\sum_{j \in [n]} |a_{ij}|^2 \leq 1 + \epsilon$$

and for every $j \in [n]$

$$\sum_{i \in [m]} |a_{ij}|^2 \geq m/n - \epsilon.$$

Proof. Let $B = (b_{ij}) = (|a_{ij}|^2)$. Then B is a real non-negative matrix satisfying Property-S. Applying Theorem 3.6 we get that for all $\epsilon > 0$ there exists a scaling $B' = \mathbf{SC}(B, \rho, \gamma)$, with ρ, γ positive real vectors, which has row sums at most $1+\epsilon$ and column sums at least $m/n-\epsilon$. Letting $\rho'_i = \sqrt{\rho_i}$ and $\gamma'_i = \sqrt{\gamma_i}$ we get a scaling $\mathbf{SC}(A, \rho', \gamma')$ of A with the required properties. □

We will use a variant of a well known lemma (see for example [Alo09]) which provides a bound on the rank of matrices whose diagonal entries are much larger than the off-diagonal ones.

Lemma 3.8. *Let $A = (a_{ij})$ be an $n \times n$ complex hermitian matrix and let $0 < \ell < L$ be integers. Suppose that $a_{ii} \geq L$ for all $i \in [n]$ and that $|a_{ij}| \leq \ell$ for all $i \neq j$. Then*

$$\mathit{rank}(A) \geq \frac{n}{1 + n \cdot (\ell/L)^2} \geq n - (n\ell/L)^2.$$

Proof. We can assume w.l.o.g. that $a_{ii} = L$ for all i. If not, then we can make the inequality into an equality by multiplying the i'th row and column by $(L/a_{ii})^{1/2} < 1$ without changing the rank or breaking the symmetry. Let $r = \mathrm{rank}(A)$ and let $\lambda_1, \ldots, \lambda_r$ denote the non-zero eigenvalues of A (counting multiplicities). Since A is hermitian we have that the λ_i's are real. We have

$$\begin{aligned} n^2 \cdot L^2 &= \mathrm{tr}(A)^2 = \left(\sum_{i=1}^{r} \lambda_i\right)^2 \leq r \cdot \sum_{i=1}^{r} \lambda_i^2 = r \cdot \sum_{i,j=1}^{n} |a_{ij}|^2 \\ &\leq r \cdot (n \cdot L^2 + n^2 \cdot \ell^2). \end{aligned}$$

Rearranging we get the required bound. The second inequality in the statement of the lemma follows from the fact that $1/(1+x) \geq 1 - x$ for all x. □

3.2 Proof of Theorem 3.2

To prove the theorem we will first find a scaling of A so that the norms (squared) of the columns are large and such that each entry is small.

Our first step is to find an $nk \times n$ matrix B that will satisfy Property-S and will be composed from rows of A s.t. each row is repeated with multiplicity between 0 and q. To achieve this we will describe an algorithm that builds the matrix B iteratively by concatenating to it rows from A. The algorithm will *mark* entries of A as it continues to add rows. Keeping track of these marks will help us decide which rows to add next. Initially all the entries of A are *unmarked*. The algorithm proceeds in k steps. At step i (i goes from 1 to k) the algorithm picks n rows from A and adds them to B. These n rows are chosen as follows: For every $j \in \{1, \ldots, n\}$ pick a row that has an unmarked non-zero entry in the j'th column and mark this non-zero entry. The reason why such a row exists at all steps is that each column contains at least k non-zero entries, and in each step we mark at most one non-zero entry in each column.

Claim 3.9. *The matrix B obtained by the algorithm has Property-S and each row of A is added to B at most q times.*

Proof. The n rows added at each of the k steps form an $n \times n$ matrix with non-zero diagonal. Thus they satisfy Property-S. It is an easy exercise to verify that a concatenation of matrices with Property-S also has this property. The bound on the number of times each row is added to B follows from the fact that each row has at most q non-zero entries and each time we add a row to B we mark one of its non-zero entries. □

Our next step is to obtain a scaling of B and, from it, a scaling of A. Fix some $\epsilon > 0$ (which will later tend to zero). Applying Corollary 3.7 we get a scaling B' of B such that the ℓ_2-norm

of each row is at most $\sqrt{1+\epsilon}$ and the ℓ_2-norm of each column is at least $\sqrt{nk/n-\epsilon} = \sqrt{k-\epsilon}$. We now obtain a scaling A' of A as follows: The scaling of the columns are the same as for B'. For the rows of A appearing in B we take the maximal scaling coefficient used for these rows in B', that is, if row i in A appears as rows $i_1, i_2, \ldots, i_{q'}$ in B, then the scaling coefficient of row i in A' is the maximal scaling coefficient of rows $i_1, i_2, \ldots, i_{q'}$ in B'. For rows *not* in B, we pick scaling coefficients so that their ℓ_2 norm (in the final scaling) is equal to 1.

Claim 3.10. *The matrix A' is a scaling of A such that each row has ℓ_2-norm at most $\sqrt{1+\epsilon}$ and each column has ℓ_2-norm at least $\sqrt{(k-\epsilon)/q}$.*

Proof. The fact that the row norms are at most $\sqrt{1+\epsilon}$ is trivial. To argue about the column norms observe that a column of B' is obtained from repeating each non-zero element in the corresponding column of A' at most q times (together with some zeros). Therefore, if we denote by $c_1, \ldots, c_s$ the non-zero entries in some column of A', we have that

$$\sum_{i=1}^{s} m_i \cdot |c_i|^2 \geq k - \epsilon,$$

where the m_i's are integers between 0 and q. In this last inequality we also relied on the fact that we chose the maximal row scaling coefficient among all those that correspond to the same row in A. Therefore,

$$\sum_{i=1}^{s} |c_i|^2 \geq (k-\epsilon)/q,$$

as required. □

Our final step is to argue about the rank of A' (which is the same as the rank of A). To this end, consider the matrix

$$M = (A')^* \cdot A',$$

where $(A')^*$ is A' transposed conjugate. Then $M = (m_{ij})$ is an $n \times n$ hermitian matrix. The diagonal entries of M are exactly the squares of the ℓ_2-norm of the columns of A'. Therefore,

$$m_{ii} \geq (k-\epsilon)/q$$

for all $i \in [n]$.

We now upper bound the off-diagonal entries. The off-diagonal entries of M are the inner products of different columns of A'. The intersection of the support of each pair of different columns is at most t. The norm of each row is at most $\sqrt{1+\epsilon}$. For every two real numbers α, β so that $\alpha^2 + \beta^2 \leq 1 + \epsilon$ we have $|\alpha \cdot \beta| \leq 1/2 + \epsilon'$, where ϵ' tends to zero as ϵ tends to zero. Therefore

$$|m_{ij}| \leq t \cdot (1/2 + \epsilon')$$

for all $i \neq j \in [n]$. Applying Lemma 3.8 we get that

$$\mathrm{rank}(A) = \mathrm{rank}(A') \geq n - \left(\frac{q \cdot t(1/2+\epsilon') \cdot n}{k-\epsilon} \right)^2.$$

Since this holds for all $\epsilon > 0$ it holds also for $\epsilon = 0$, which gives the required bound on the rank of A. □

4 Sylvester-Gallai Configurations

In this section we prove the quantitative Sylvester-Gallai (SG) Theorem. We will be interested with point configurations in real and complex space. These are finite sets of distinct points $v_1, \ldots, v_n$ in $\mathbb{R}^d$ or $\mathbb{C}^d$. The dimension of a configuration is defined to be the dimension of the smallest affine subspace containing all points.

Definition 4.1 (Special and ordinary lines)**.** Let $v_1, \ldots, v_n \in \mathbb{C}^d$ be a set of n distinct points in d-dimensional complex space. A line ℓ passing through at least three of these points is called a *special* line. A line passing through exactly two points is called an *ordinary* line.

Definition 4.2 (δ-SG configuration)**.** Let $\delta \in [0, 1]$. A set of n distinct points $v_1, \ldots, v_n \in \mathbb{C}^d$ is called a *δ-SG configuration* if for every $i \in [n]$, there exists a family of special lines L_i all passing through v_i and at least δn of the points $v_1, \ldots, v_n$ are on the lines in L_i. (Note that each collection L_i may cover a different subset of the n points.)

The main result of this section bounds the dimension of δ-SG configurations for all $\delta > 0$. Since we can always satisfy the definition by spreading the points evenly over $1/\delta$ lines we know that the dimension can be at least $2/\delta$ (and in fact in complex space at least $3/\delta$). We prove an upper bound of $O(1/\delta^2)$.

Theorem 4.3 (Restatement of Theorem 2 – quantitative SG theorem)**.** *Let $\delta \in (0, 1]$. Let $v_1, \ldots, v_n \in \mathbb{C}^d$ be a δ-SG configuration. Then*

$$dim\{v_1, \ldots, v_n\} < 13/\delta^2.$$

Moreover, the dimension of a 1-SG configuration is at most 10.

The constants in the proof have been optimized to the best of our abilities. Notice that in the above theorem δ can be dependant on n. For example, a $(1/\log(n))$-SG configuration of n points can have rank at most $O(\log(n)^2)$.

4.1 Preliminaries to the Proof of Theorem 4.3

The notion of a latin square will turn out useful in the proof:

Definition 4.4 (Latin squares)**.** An *$r \times r$ latin square* is an $r \times r$ matrix D such that $D_{i,j} \in [r]$ for all i, j and every number in $[r]$ appears exactly once in each row and in each column. A latin square D is called *diagonal* if $D_{i,i} = i$ for all $i \in [r]$.

Theorem 4.5 ([Hil73])**.** *For every $r \geq 3$ there exists a diagonal $r \times r$ latin square.*

We note that we use diagonal latin squares only to optimize constant factors. If one does not care about such factors then there is a simple construction that serves the same goal.

The following lemma is an easy consequence of the above theorem.

Lemma 4.6. *Let $r \geq 3$. Then there exists a set $T \subset [r]^3$ of $r^2 - r$ triples that satisfies the following properties:*

1. *Each triple* $(t_1, t_2, t_3) \in T$ *is of three distinct elements.*

2. *For each* $i \in [r]$ *there are exactly* $3(r-1)$ *triples in* T *containing* i *as an element.*

3. *For every pair* $i, j \in [r]$ *of distinct elements there are at most* 6 *triples in* T *which contain both* i *and* j *as elements.*

Proof. Let D be an $r \times r$ diagonal latin square which we know exists from Theorem 4.5. Define $T \subset [r]^3$ to be the set of all triples $(i, j, k) \in [r]^3$ with $i \neq j$ such that $D_{i,j} = k$. The number of such triples is $r^2 - r$. Property 1 holds by the definition of diagonal latin square— we cannot have $D_{i,j} = i$ for $j \neq i$ since $D_{i,i} = i$ and every row in D has distinct as the (i, i) entry in D is labeled i for all $i \in [r]$, and similarly we cannot have $D_{i,j} = j$ for $i \neq j$.

Let $i \in [r]$. By construction, there are $r-1$ triples in T which have i as their first entry, and $r-1$ triples that have i as their second entry. There are also $r-1$ triples in T which have i as their last entry, since for every one of the $r-1$ rows $i' \neq i$ there is exactly one location $j' \neq i'$ in which the label i appears, and that contributes the triple (i', j', i) to T. This proves Property 2.

To prove Property 3 observe that two triples in T can agree in at most one place. For example, knowing the row and column determines the label, knowing the row and label determines the column, and so forth. Therefore, a pair (i, j) cannot appear in more than 6 triples since otherwise there would have been at least two triples with i, j at the same places, and these triples would violate the above rule. □

4.2 Proof of Theorem 4.3

Let V be the $n \times d$ matrix whose i'th row is the vector v_i. Assume w.l.o.g. that $v_1 = 0$. Thus

$$\dim\{v_1, \ldots, v_n\} = \text{rank}(V).$$

The overview of the proof is as follows. We will first build an $m \times n$ matrix A that will satisfy $A \cdot V = 0$. Then, we will argue that the rank of A is large because it is a design matrix. This will show that the rank of V is small.

Consider a special line ℓ which passes through three points v_i, v_j, v_k. This gives a linear dependency among the three vectors v_i, v_j, v_k (we identify a point with its vector of coordinates in the standard basis). In other words, this gives a vector $a = (a_1, \ldots, a_n)$ which is non-zero only in the three coordinates i, j, k and such that $a \cdot V = 0$. If a is not unique, choose an arbitrary vector a with these properties.

Our strategy is to pick a family of collinear triples among the points in our configuration and to build the matrix A from rows corresponding to these triples in the above manner.

Let $\mathcal{L}$ denote the set of all special lines in the configuration (i.e. all lines containing at least three points). Then each L_i is a subset of $\mathcal{L}$ containing lines passing through v_i. For each $\ell \in \mathcal{L}$ let V_ℓ denote the set of points in the configuration which lie on the line ℓ. Then $|V_\ell| \geq 3$ and we can assign to it a family of triples $T_\ell \subset V_\ell^3$, given by Lemma 4.6 (we identify V_ℓ with $[r]$, where $r = |V_\ell|$ in some arbitrary way).

We now construct the matrix A by going over all lines $\ell \in \mathcal{L}$ and for each triple in T_ℓ adding as a row of A the vector with three non-zero coefficients $a = (a_1, \ldots, a_n)$ described above (so that a is the linear dependency between the three points in the triple).

Since the matrix A satisfies $A \cdot V = 0$ by construction, we only have to argue that A is a design matrix and bound its rank.

Claim 4.7. *The matrix A is a $(3, 3k, 6)$-design matrix, where $k \triangleq \lfloor \delta n \rfloor - 1$.*

Proof. By construction, each row of A has exactly 3 non-zero entries. The number of non-zero entries in column i of A corresponds to the number of triples we used that contain the point v_i. These can come from all special lines containing v_i. Suppose there are s special lines containing v_i and let $r_1, \ldots, r_s$ denote the number of points on each of those lines. Then, since the lines through v_i have only the point v_i in common, we have that

$$\sum_{j=1}^{s} (r_j - 1) \geq k.$$

The properties of the families of triples T_ℓ guarantee that there are $3(r_j - 1)$ triples containing v_i coming from the j'th line. Therefore there are at least $3k$ triples in total containing v_i.

The size of the intersection of columns i_1 and i_2 is equal to the number of triples containing the points v_{i_1}, v_{i_2} that were used in the construction of A. These triples can only come from one special line (the line containing these two points) and so, by Lemma 4.6, there can be at most 6 of those. □

Applying Theorem 3.2 we get that

$$\begin{aligned} \operatorname{rank}(A) &\geq n - \left(\frac{3 \cdot 6 \cdot n}{2 \cdot 3k}\right)^2 \geq n - \left(\frac{3 \cdot n}{\delta n - 2}\right)^2 \\ &\geq n - \left(\frac{3 \cdot n \cdot 13}{11 \cdot \delta n}\right)^2 > n - 13/\delta^2, \end{aligned}$$

where the third inequality holds as $\delta n \geq 13$ since otherwise the theorem trivially holds. Since $A \cdot V = 0$ we have that

$$\operatorname{rank}(A) + \operatorname{rank}(V) \leq n.$$

This implies that

$$\operatorname{rank}(V) < 13/\delta^2,$$

which completes the proof. For $\delta = 1$, the calculation above yields $\operatorname{rank}(V) < 11$. □

4.3 Average-Case Version

In this section we use Theorem 4.3 to argue about the case where we only know that there are many collinear triples in a configuration.

Theorem 4.8 (Average-case SG theorem)**.** *Let $V = \{v_1, \ldots, v_m\} \subset \mathbb{C}^d$ be a set of m distinct points. Let T be the set of (unordered) collinear triples in V. Suppose $|T| \geq \alpha m^2$ and that every two points v, v' in V appear in at most c triples in T, then there exists a subset $V' \subset V$ such that $|V'| \geq \alpha m/(2c)$ and $dim(V') \leq O(1/\alpha^2)$.*

Notice that the bound on the number of triples containing a fixed pair of points is necessary for the theorem to hold. If we remove this assumption than we could create a counter-example by arranging the points so that $m^{2/3}$ of them are on a line and the rest span the entire space.

Lemma 4.9. *Let H be a 3-regular hypergraph with vertex set $[m]$ and αm^2 edges of co-degree at most c (i.e. for every $i \neq j$ in $[m]$, the set $\{i, j\}$ is contained in at most c edges). Then there is a subset $M \subseteq [m]$ of size $|M| \geq \alpha m/(2c)$ so that the minimal degree of the sub-graph of H induced by M is at least $\alpha m/2$.*

Proof. We describe an iterative process to find M. We start with $M = [m]$. While there exists a vertex of degree less than $\alpha m/2$, remove this vertex from M and remove all edges containing this vertex from H. Continuing in this fashion we conclude with a set M such that every point in M has degree at least $\alpha m/2$. This process removed in total at most $m \cdot \alpha m/2$ edges and thus the new H still contains at least $\alpha m^2/2$ edges. As the co-degree is at most c, every vertex appears in at most cm edges. Thus, the size of M is of size at least $\alpha m/(2c)$. □

Proof of Theorem 4.8. The family of triples T defines a 3-regular hypergraph on V of co-degree at most c. Lemma 4.9 thus implies that there is a subset $V' \subseteq V$ of size $|V'| \geq \alpha m/(2c)$ that is an $(\alpha/2)$-SG configuration. By Theorem 4.3, V' has dimension at most $O(1/\alpha^2)$. □

5 Robust SG Theorem for k-Flats

In this section we prove two high-dimensional analogs of the SG theorem. Let $\mathrm{fl}(v_1, \ldots, v_k)$ (fl for 'flat') denote the affine span of k points (i.e. the points that can be written as linear combinations with coefficients that sum to one). We call $v_1, \ldots, v_k$ *independent* if their flat is of dimension $k-1$ (dimension means affine dimension), and say that $v_1, \ldots, v_k$ are *dependent* otherwise. A *k-flat* is an affine subspace of dimension k.

In the following V is a set of n distinct points in complex space $\mathbb{C}^d$. A k-flat is called *ordinary* if its intersection with V is contained in the union of a $(k-1)$-flat and a single point. A k-flat is *elementary* if its intersection with V has exactly $k+1$ points. Notice that for $k = 1$ (lines) the two notions of ordinary and elementary coincide.

For dimensions higher than one, there are two different definitions that generalize that of SG configuration. The first definition is based on ordinary k-flats (though in a slightly stronger way which will be more useful in the proofs to come). The second definition (which is less restricted than the first one) uses elementary k-flats.

Definition 5.1. The set V is a δ-SG_k^* configuration if for every independent $v_1, \ldots, v_k \in V$ there are at least δn points $u \in V$ s.t. either $u \in \mathrm{fl}(v_1, \ldots, v_k)$ or the k-flat $\mathrm{fl}(v_1, \ldots, v_k, u)$ contains a point w outside $\mathrm{fl}(v_1, \ldots, v_k) \cup \{u\}$.

Definition 5.2. The set V is a δ-SG$_k$ configuration if for every independent $v_1, \ldots, v_k \in V$ there are at least δn points $u \in V$ s.t. either $u \in \mathrm{fl}(v_1, \ldots, v_k)$ or the k-flat $\mathrm{fl}(v_1, \ldots, v_k, u)$ is not elementary.

Both definitions coincide with that of SG configuration when $k = 1$: Indeed, $\mathrm{fl}(v_1) = v_1$ and $\mathrm{fl}(v_1, u)$ is the line through v_1, u. Therefore, u is never in $\mathrm{fl}(v_1)$ and the line $\mathrm{fl}(v_1, u)$ is not elementary iff it contains at least one point $w \notin \{v_1, u\}$.

We prove two high-dimensional versions of the SG theorem, each corresponding to one of the definitions above. The first uses the more restricted 'star' definition and gives a strong upper bound on dimension. The second uses the less restricted definition and gives a weaker bound on dimension.

Theorem 5.3. *Let V be a δ-SG$^*_k$ configuration. Then $dim(V) \leq f(\delta, k)$ with*

$$f(\delta, k) = O\left((k/\delta)^2\right).$$

Theorem 5.4. *Let V be a δ-SG$_k$ configuration. Then $dim(V) \leq g(\delta, k)$ with*

$$g(\delta, k) = 2^{C^k}/\delta^2$$

with $C > 1$ a universal constant.

The proofs of the two theorems are below. Theorem 5.3 follows by an appropriate induction on the dimension, using the (one-dimensional) robust SG theorem. Theorem 5.4 follows by reduction to Theorem 5.3.

Before proving the theorems we set some notations. Fix some point $v_0 \in V$. By a *normalization w.r.t.* v_0 we mean an affine transformation $N : \mathbb{C}^d \mapsto \mathbb{C}^d$ which first moves v_0 to zero, then picks a hyperplane H s.t. no point in V (after the shift) is parallel to H (i.e has inner product zero with the orthogonal vector to H) and finally multiplies each point (other than zero) by a constant s.t. it is in H.

Claim 5.5. *For such a mapping N we have that $v_0, v_1, \ldots, v_k$ are dependent iff $N(v_1), \ldots, N(v_k)$ are dependent.*

Proof. Since translation and scaling does not affect dependence, w.l.o.g. we assume that $v_0 = 0$ and that the distance of the hyperplane H from zero is one. Let h be the unit vector orthogonal to H. For all $i \in [k]$ we have $N(v_i) = v_i/\langle v_i, h\rangle$. Assume that $v_0, v_1, \ldots, v_k$ are dependent, that is, w.l.o.g. $v_k = \sum_{i\in[k-1]} a_i v_i$ for some $a_1, \ldots, a_{k-1}$. For all $i \in [k-1]$ define $b_i = a_i\langle v_i, h\rangle/\langle v_k, h\rangle$. Thus $N(v_k) = \sum_{i\in[k-1]} a_i v_i/\langle v_k, h\rangle = \sum_{i\in[k-1]} b_i N(v_i)$ where $\sum_{i\in[k-1]} b_i = 1$, which means that $N(v_1), \ldots, N(v_k)$ are dependent. Since the map $a_i \mapsto b_i$ is invertible, the other direction of the claim holds as well. □

We first prove the theorem for δ-SG$^*_k$ configurations.

Proof of Theorem 5.3. The proof is by induction on k. For $k = 1$ we know $f(\delta, 1) \leq c\delta^{-2}$ with $c > 1$ a universal constant. Suppose $k > 1$. We separate into two cases. The first case is when

V is an $(\delta/(2k))$-SG$_1$ configuration and we are done using the bound on $k = 1$. In the other case there is some point $v_0 \in V$ s.t. the size of the set of points on special lines through v_0 is at most $\delta/(2k)$ (a line is special if it contains at least three points). Let S denote the set of points on special lines through v_0. Thus $|S| < \delta n/(2k)$. Let $N : \mathbb{C}^d \mapsto \mathbb{C}^d$ be a normalization w.r.t. v_0. Notice that for points $v \notin S$ the image $N(v)$ determines v. Similarly, all points on some special line map to the same point via N.

Our goal is to show that $V' = N(V \setminus \{v_0\})$ is a $((1 - 1/(2k))\delta)$-SG^*_{k-1} configuration (after eliminating multiplicities from V'). This will complete the proof since $\dim(V) \leq \dim(V') + 1$. Indeed, if this is the case we have

$$f(\delta, k) \leq \max\{4c(k/\delta)^2, f((1 - 1/(2k))\delta, k - 1) + 1\}.$$

and by induction we have $f(\delta, k) \leq 4c(k/\delta)^2$.

Fix $v'_1, \ldots, v'_{k-1} \in V'$ to be $k-1$ independent points (if no such tuple exists then V' is trivially a configuration). Let $v_1, \ldots, v_{k-1} \in V$ be points s.t. $N(v_i) = v'_i$ for $i \in [k-1]$. Claim 5.5 implies that $v_0, v_1, \ldots, v_{k-1}$ are independent. Thus, there is a set $U \subset V$ of size at least δn s.t. for every $u \in U$ either $u \in \text{fl}(v_0, v_1, \ldots, v_{k-1})$ or the k-flat $\text{fl}(v_0, v_1, \ldots, v_{k-1}, u)$ contains a point w outside $\text{fl}(v_0, v_1, \ldots, v_{k-1}) \cup \{u\}$.

Let $\tilde{U} = U \setminus S$ so that N is invertible on $\tilde{U}$ and

$$|\tilde{U}| \geq |U| - |S| \geq (1 - 1/(2k))\delta n.$$

Suppose $u \in \tilde{U}$ and let $u' = N(u)$. By Claim 5.5 if $u \in \text{fl}(v_0, v_1, \ldots, v_{k-1})$ then u' is in $\text{fl}(v'_1, \ldots, v'_{k-1})$. Otherwise, $\text{fl}(v_0, v_1, \ldots, v_{k-1}, u)$ contains a point w outside $\text{fl}(v_0, v_1, \ldots, v_{k-1}) \cup \{u\}$. Let $w' = N(w)$. We will show that w' is (a) contained in the $(k-1)$-flat $\text{fl}(v'_1, \ldots, v'_{k-1}, u')$ and (b) is outside $\text{fl}(v'_1, \ldots, v'_{k-1}) \cup \{u'\}$. Property (a) follows from Claim 5.5 since $v_0, v_1, \ldots, v_{k-1}$, u, w are dependent and so $v'_1, \ldots, v'_{k-1}, u', w'$ are also dependent. To show (b) observe first that by Claim 5.5 the points $v'_1, \ldots, v'_{k-1}, u'$ are independent (since $v_0, v_1, \ldots, v_{k-1}, u$ are independent) and so u' is not in $\text{fl}(v'_1, \ldots, v'_{k-1})$. We also need to show that $w' \neq u'$ but this follows from the fact that $u \neq w$ and so $w' = N(w) \neq N(u) = u'$ since N is invertible on $\tilde{U}$ and $u \in \tilde{U}$. Since

$$|N(\tilde{U})| = |\tilde{U}| \geq (1 - 1/(2k))\delta n \geq (1 - 1/(2k))\delta |V'|$$

the proof is complete. □

We can now prove the theorem for δ-SG$_k$ configurations.

Proof of Theorem 5.4. The proof follows by induction on k (the case $k = 1$ is given by Theorem 4.3). Suppose $k > 1$. Suppose that $\dim(V) > g(\delta, k)$. We want to show that there exist k independent points $v_1, \ldots, v_k$ s.t. for at least $1 - \delta$ fraction of the points $w \in V$ we have that w is not in $\text{fl}(v_1, \ldots, v_k)$ **and** the flat $\text{fl}(v_1, \ldots, v_k, w)$ is elementary (i.e. does not contain any other point).

Let $k' = g(1, k - 1)$. By choice of g we have $g(\delta, k) > f(\delta, k' + 1)$ with f from Theorem 5.3. Thus, by Theorem 5.3, we can find $k' + 1$ independent points $v_1, \ldots, v_{k'+1}$ s.t. there is a set $U \subset V$ of size at least $(1 - \delta)n$ s.t. for every $u \in U$ we have that u is not in $\text{fl}(v_1, \ldots, v_{k'+1})$ **and** the $(k' + 1)$-flat $\text{fl}(v_1, \ldots, v_{k'+1}, u)$ contains only one point, namely u, outside $\text{fl}(v_1, \ldots, v_{k'+1})$.

We now apply the inductive hypothesis on the set $V \cap \mathrm{fl}(v_1, \ldots, v_{k'+1})$ which has dimension at least $k' = g(1, k-1)$. This gives us k independent points $v'_1, \ldots, v'_k$ that define an elementary $(k-1)$-flat $\mathrm{fl}(v'_1, \ldots, v'_k)$. (Saying that V is not 1-SG$_{k-1}$ is the same as saying that it contains an elementary $(k-1)$-flat). Joining any of the points $u \in U$ to $v'_1, \ldots, v'_k$ gives us an elementary k-flat and so the theorem is proved. □

6 Generalizations of the Motzkin-Rabin Theorem

In this section we prove two variants of the Motzkin-Rabin Theorem. The first is a quantitative analog in the spirit of Theorem 4.3. The second is a variant in which the number of colors is three (instead of two).

6.1 A Quantitative Variant

Definition 6.1 (δ-MR configuration)**.** Let V_1, V_2 be two disjoint finite subsets of $\mathbb{C}^d$. Points in V_1 are of *color* 1 and points in V_2 are of *color* 2. A line is called *bi-chromatic* if it contains at least one point from each of the two colors. We say that V_1, V_2 are a *δ-MR* configuration if for every $i \in [2]$ and for every point $p \in V_i$, the bi-chromatic lines through p contain at least $\delta|V_i|$ points.

Theorem 6.2. *Let $V_1, V_2 \subset \mathbb{C}^d$ be a δ-MR configuration. Then*

$$\dim(V_1, V_2) \leq O(1/\delta^4).$$

Proof. We will call a line passing through exactly two points in V_1 (resp. V_2) a V_1-ordinary (resp. V_2-ordinary) line. W.l.o.g. assume

$$|V_1| \leq |V_2|.$$

We seperate the proof into two cases:

Case I is when V_2 is a $(\delta/2)$-SG configuration. Then, by Theorem 4.3, $\dim(V_2) \leq O(1/\delta^2)$. If in addition

$$\dim(V_1) \leq 13/(\delta/2)^2$$

then we are done. Otherwise, by Theorem 4.3, there exists a point $a_0 \in V_1$ such that there are at least $(1-\delta/2)|V_1|$ V_1-ordinary lines through a_0. Let $a_1, \ldots, a_k$ denote the points in V_1 that belong to these lines with $k \geq (1-\delta/2)|V_1|$. We now claim that $V_2 \cup \{a_0\}$ spans all the points in V_1. This will suffice since, in this case, $\dim(V_2) \leq O(1/\delta^2)$. Let $a \in V_1$. Then, since V_1, V_2 is a δ-MR configuration, there are at least $\delta|V_1|$ points in V_1 such that the line through them and a contains a point in V_2. One of these points must be among $a_1, \ldots, a_k$, say it is a_1. Since a is in the span of V_2 and a_1 and since a_1 is in the span of V_2 and a_0 we are done.

Case II is when V_2 is not a $(\delta/2)$-SG configuration. In this case, there is a point $b \in V_2$ such that there are at least $(1-\delta/2)|V_2|$ V_2-ordinary lines through b. From this fact and from the δ-MR property, we get that $|V_1| \geq (\delta/2)|V_2|$ (there are at least $(\delta/2)|V_2|$ V_2-ordinary lines through b that have an additional point from V_1 on them). This implies that the union $V_1 \cup V_2$ is a $(\delta^2/4)$-SG configuration and the result follows by applying Theorem 4.3. □

6.2 A Three Colors Variant

Definition 6.3 (3MR configuration)**.** Let V_1, V_2, V_3 be three pairwise disjoint finite subsets of $\mathbb{C}^d$, each of distinct points. We say that V_1, V_2, V_3 is a *3MR*-configuration if every line ℓ so that $\ell \cap (V_1 \cup V_2 \cup V_3)$ has more than one point intersects at least two of the sets V_1, V_2, V_3.

Theorem 6.4. *Let V_1, V_2, V_3 be a 3MR configuration and denote $V = V_1 \cup V_2 \cup V_3$. Then*

$$dim(V) \leq O(1).$$

Proof. Assume w.l.o.g. that V_1 is not smaller than V_2, V_3. Let $\alpha = 1/16$. There are several cases to consider:

1. V_1 is an α-SG configuration. By Theorem 4.3, the dimension of V_1 is at most

$$d_1 = O(1/\alpha^2).$$

Consider the two sets

$$V_2' = V_2 \setminus \text{span}(V_1) \text{ and } V_3' = V_3 \setminus \text{span}(V_1),$$

each is a set of distinct points in $\mathbb{C}^d$. Assume w.l.o.g. that $|V_2'| \geq |V_3'|$.

1.1. V_2' is an α-SG configuration. By Theorem 4.3, the dimension of V_2' is at most

$$d_2 = O(1/\alpha^2).$$

Fix a point v_3 in V_3'. For every point $v \neq v_3$ in V_3' the line through v_3, v contains a point from $\text{span}(V_1) \cup V_2'$. Therefore,

$$\dim(V) \leq d_1 + d_2 + 1 \leq O(1).$$

1.2. V_2' is not an α-SG configuration. There is a point v_2 in V_2' so that for $k \geq |V_2'|/2$ of the points $v \neq v_2$ in V_2' the line through v_2, v does not contain any other point from V_2'. If $V_2' = \text{span}(V_1, v_2)$ then the dimension of $V_1 \cup V_2$ is at most $d_1 + 1$ and we are done as in the previous case. Otherwise, there is a point v_2' in $V_2' \setminus \text{span}(V_1, v_2)$.

We claim that in this case $|V_3'| \geq k/2$. Denote by P_2 the k points $v \neq v_2$ in V_2' so that the line through v_2, v does not contain any other point from V_2'. For every $v \in P_2$ there is a point $V_{1,3}(v)$ in $V_1 \cup V_3$ that is on the line through v, v_2 (the point v_2 is fixed). There are two cases to consider.

The first case is that for at least $k/2$ of the points v in P_2 we have $V_{1,3}(v) \in V_3$. In this case clearly $|V_3| \geq k/2$.

The second case is that for at least $k/2$ of the points v in P_2 we have $V_{1,3}(v) \in V_1$. Fix such a point $v \in P_2$ (which is in $\text{span}(V_1, v_2)$). The line through v_2', v contains a point v' from $V_1 \cup V_3$. The point v' is not in $\text{span}(V_1)$, as if it was then v_2' would be in $\text{span}(v, v') \subseteq \text{span}(V_1, v)$. Therefore v' is in V_3. This also implies that $|V_3'| \geq k/2$.

Denote $V' = V_2 \cup V_3'$. So we can conclude that for every v' in V' the special lines through v' contain at least $|V'|/8$ of the points in $V_1 \cup V_2 \cup V_3$. As in the proof of Theorem 4.3, we can thus define a family of triples T, each triple of three distinct collinear points in V, so that each v' in V' belongs to at least $|V'|/8$ triples in T and each two distinct v', v'' in V' belong to at most 6 triples.

By a slight abuse of notation, we also denote by V the matrix with rows defined by the points in V. Let V_1 be the submatrix of V with row defined by points in $\text{span}(V_1) \cap V$ and V' be the submatrix of V with row defined by points in V'. Use the triples in T to construct a matrix A so that $A \cdot V = 0$. Let A_1 be the submatrix of A consisting of the columns that correspond to $\text{span}(V_1) \cap V$ and A' be the submatrix of A consisting of the columns that correspond to V'. Therefore, $A' \cdot V' = -A_1 \cdot V_1$ which implies

$$\text{rank}(A' \cdot V') \leq \text{rank}(A_1 \cdot V_1) \leq d_1.$$

By the above discussion A' is a $(3, |V'|/8, 6)$-design matrix and thus, by Theorem 3.2, has rank at least

$$|V'| - O(1)$$

and so

$$\dim(V') \leq O(1) + d_1 \leq O(1).$$

We can finally conclude that

$$\dim(V) \leq d_1 + \dim(V') \leq O(1).$$

2. V_1 **is not an** α**-SG configuration.** There is a point v_1 in V_1 so that for at least $|V_1|/2$ of the points $v \neq v_1$ in V_1 the line through v_1, v does not contain any other point from V_1. Assume w.l.o.g. that $|V_2| \geq |V_3|$. This implies that

$$|V_2| \geq |V_1|/4.$$

2.1. $|V_3| < |V_2|/16$**.** In this case the configuration defined by $V_1 \cup V_2$ is an α-SG configuration. By Theorem 4.3, the dimension of $V_1 \cup V_2$ is at most

$$d_{1,2} = O(1/\alpha^2).$$

Fix a point v_3 in V_3. For every point $v \neq v_3$ in V_3 the line through v_3, v contains a point from $V_1 \cup V_2$. Therefore,

$$\dim(V) \leq d_{1,2} + 1 \leq O(1).$$

2.1. $|V_3| \geq |V_2|/16$**.** In this case V is an α-SG configuration. By Theorem 4.3, the dimension of V is thus at most $O(1/\alpha^2)$.

□

7 Two-Query Locally Correctable Codes

We now prove the non-existence of 2-query (linear) locally correctable codes (LCC) over $\mathbb{C}$. We start by formally defining locally correctable codes:

Definition 7.1 (Linear locally correctable code (LCC))**.** Let $\mathbb{F}$ be some field. A (q, δ)-LCC over $\mathbb{F}$ is a linear subspace $C \subset \mathbb{F}^m$ such that there exists a randomized decoding procedure $D : \mathbb{F}^m \times [m] \mapsto \mathbb{F}$ with the following properties:

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

The *dimension* of an LCC is simply its dimension as a subspace of $\mathbb{F}^m$.

In the above definition we allow the algorithm D to perform operations over the field $\mathbb{F}$. Since we do not care about the running time of D we do not discuss issues of representation of field elements and efficiency of handling them. (In any case, it turns out that for linear codes in the small number of queries and low error case, one can assume w.l.o.g. that the decoder is also linear, see Lemma 7.4 below.)

Our result on locally decodable codes is the following:

Theorem 7.2 (Restatement of Theorem 3— non-existence of 2 query LCCs over $\mathbb{C}$)**.** *Let $C \subset \mathbb{C}^m$ be a $(2, \delta)$-LCC over $\mathbb{C}$. Then*

$$\dim(C) \leq O(1/\delta^9).$$

As in Theorem 4.3, also in this theorem, δ can be an arbitrary function of m. To make the connection between LCCs and SG-configurations explicit, we define the notion of a δ-LCC configuration.

Definition 7.3 (δ-LCC Configuration)**.** A list of non-zero points $(v_1, \ldots, v_m)$ in $\mathbb{C}^d$ (not necessarily distinct) is called a δ-LCC configuration if for every subset $\Delta \subset [m]$ of size at most δm and for every $i \in [m]$, there exist $j, k \in [m] \setminus \Delta$ such that either $v_i \in \{v_j, v_k\}$ (in which case v_i can be recovered by its own copies), or v_i, v_j, v_k are three distinct collinear points (in which case v_i is recovered by two other coordinates).

The following lemma shows the connection between these two notions.

Lemma 7.4. *If there exists a $(2, \delta)$-LCC of dimension n over $\mathbb{C}$ then there exists a δ-LCC configuration of dimension at least $n - 1$ over $\mathbb{C}$.*

To prove the lemma we will use the following definition.

Definition 7.5 (Generating set)**.** Let $C \subset \mathbb{F}^m$ be a subspace. We say that a list of vectors $V = (v_1, \ldots, v_m)$ in $\mathbb{F}^n$ is a *generating set* for C if

$$C = \{(\langle y, v_1\rangle, \langle y, v_2\rangle, \ldots, \langle y, v_m\rangle) \mid y \in \mathbb{F}^n\},$$

where $\langle y, v\rangle$ is the standard inner product over $\mathbb{F}$.

Proof of Lemma 7.4. Let $V = (v_1, \ldots, v_m)$ be a generating set for C with $\dim(V) \geq n-1$. We might lose 1 since we defined $\dim(V)$ as the dimension of the smallest *affine* subspace containing V. When the local decoder for C reads two positions in a codeword, it is actually reading $\langle y, v_j\rangle, \langle y, v_k\rangle$ for some vector $y \in \mathbb{C}^n$ (or noisy versions of them). In order to be able to recover $\langle y, v_i\rangle$ from $\langle y, v_j\rangle, \langle y, v_k\rangle$ with positive probability it must be that $v_i \in \mathrm{span}\{v_j, v_k\}$. (If we choose y as Gaussian and v_i is not in the span of v_j, v_k then even conditioned on the values of $\langle y, v_j\rangle, \langle y, v_k\rangle$ the r.v. $\langle y, v_i\rangle$ takes any specific value with probability zero.) Applying an invertible linear transformation on V preserves properties such as one vector being in the span of another set. So we can assume w.l.o.g. that the first coordinate in all elements of V is non-zero. Scaling each v_i by a non-zero scalar also preserves the properties of spans and so we can assume w.l.o.g. that the first coordinate in each v_i is equal to 1. Now, for v_i to be in the span of v_j, v_k it must be that either $v_i \in \{v_j, v_k\}$ or v_i is on the line passing through v_j, v_k (and they are all distinct). Thus, we have a δ-LCC configuration with dimension $n-1$. □

In view of this lemma, in order to prove Theorem 7.2 it is enough to prove:

Theorem 7.6. *Let $V = (v_1, \ldots, v_m) \in (\mathbb{C}^d)^m$ be a δ-LCC configuration. Then*

$$dim(V) \leq O(1/\delta^9).$$

7.1 Proof of Theorem 7.6

Let $V = (v_1, \ldots, v_m)$ be the list of m points in $\mathbb{C}^d$. The main difficulty in proving the theorem is that some of these points may be the same. That is, two points v_i, v_j can actually correspond to the same vector in $\mathbb{C}^d$. In this case we say that v_i, v_j are *copies* of each other. Otherwise, we say that v_i, v_j are *distinct*. If v is a point in the list V, we let the *multiplicity* of v, denoted $M(v)$, be the number of times that (a copy of) v occurs in V.

We note that while repetitions make the proof of Theorem 7.6 more complicated, we do not know if they actually help in constructing LCCs with better parameters. Our proof will proceed in an iterative way, at each step identifying a sufficiently large sublist with small dimension and removing it. The key step will be the following theorem:

Theorem 7.7. *There exists an integer $K_1 > 0$ s.t. the following holds. Let $V = (v_1, \ldots, v_m) \in (\mathbb{C}^d)^m$ be a δ-LCC configuration. Then there exists a sublist $V' \subset V$ of size at least $\delta^3 m/K_1$ and dimension at most K_1/δ^6.*

Proof. If there exists a point $v \in V$ with multiplicity larger that $\delta m/10$ then the theorem is true by taking V' to be all copies of this point. This avoids the case where a point is recovered mostly by its own copies. For the rest of the proof we can, thus, assume the following.

Fact 7.8. *For all $v \in V$ and for every sublist Δ of V of size at most $\delta m/2$ there is a collinear triple containing v such that the other two points in the triple are not in Δ (and are distinct from v).*

We will describe a (probabilistic) construction of a family of collinear triples and build a design matrix from it. We call a triple of points in V *good* if it contains three distinct collinear points. We define a family T of good triples as follows: For every line ℓ that has at least three distinct points in V we will define (randomly) a family T_ℓ of good triples (later we will fix the randomness). The family T will be the union of all these sets.

Remark 7.9. The construction of T we present is probabilistic. It is possible to construct T explicitly and achieve similar properties. We choose to present the probabilistic construction as it is simpler and less technical.

Let ℓ be such a line with r points on it (counting multiplicities). Denote by $V(\ell)$ the sublist of V containing all points that lie on ℓ. We first take the family F of triples on $[r]$ given by Lemma 4.6 and then pick a random one-to-one mapping $\rho : [r] \mapsto V(\ell)$. For a triple t in F we denote by $\rho(t)$ the triple of points in $V(\ell)$ that is the image of t under ρ. We take T_ℓ to be the set of all triples $\rho(t)$ with $t \in F$ and such that $\rho(t)$ is good (i.e., it 'hits' three distinct points).

Intuitively, we will have many good triples on a line (in expectation) if there are no two points whose copies cover most of the line (then the probability of hitting three distinct points is small). We will later show that this cannot happen on too many lines.

The next proposition shows that there is a way to fix the randomness so that T contains a quadratic number of triples.

Proposition 7.10. *The expectation of $|T|$ is at least αm^2 with $\alpha = (\delta/15)^3$.*

We will prove this proposition later in Section 7.2 and will continue now with the proof of the theorem.

Fix T to be a family of triples that has size at least the expectation of $|T|$. By construction and Lemma 4.6, the family T contains only good triples and each pair of points appears in at most 6 different triples (since every two distinct points define a single line and two non-distinct points never appear in a triple together). The family T thus defines a 3-regular hypergraph with vertex set $[m]$ and at least αm^2 edges and of co-degree at most 6. Lemma 4.9 thus implies that there is a sublist V' of V of size at least

$$|V'| = m' \geq \alpha m/12 \geq (\delta/45)^3 m$$

with the following property: Let T' be the subfamily of T that V' induces. Every v' in V' is contained in at least $\alpha m/2$ triples in T'.

By a slight abuse of notation, we also denote by V' the $m' \times d$ matrix with rows defined by the points in V' (including repetitions). We now use the triples in T' to construct a matrix A' so that $A' \cdot V' = 0$. By the above discussion A' is a $(3, \alpha m/2, 6)$-design matrix and thus, by Theorem 3.2, has rank at least

$$m' - \left(\frac{18m'}{\alpha m}\right)^2 \geq m' - (18/\alpha)^2$$

and so

$$\dim(V') \leq (18/\alpha)^2 \leq (60/\delta)^6$$

as was required. □

The next proposition shows how the above theorem can be used repeatedly on a given LCC.

Proposition 7.11. *There exist an integer $K_2 > 0$ s.t. the following holds: Let $V = (v_1, \ldots, v_m) \in (\mathbb{C}^d)^m$ be a δ-LCC configuration and let U, W be a partition of V into two disjoint sublists such that $W \cap span(U) = \emptyset$. Then there exists a new partition of V to two sublists U' and W' such that $W' \cap span(U') = \emptyset$ and such that*

1. *$|U'| \geq |U| + \delta^3 m/K_2$, and*
2. *$dim(U') \leq dim(U) + K_2/\delta^6$.*

Proof. First, we can assume that all points in W have multiplicity at most $\delta m/2$ (otherwise we can add one point from W with high multiplicity to U to get U'). Thus, for all points v and all sublists Δ of size at most $\delta m/2$ there is a collinear triple of three distinct points containing v and two other points outside Δ. Again, this is to avoid points that are recovered mostly by copies of themselves.

For a point $w \in W$ we define three disjoint sublists of points $U(w), P_1(w)$ and $P_2(w)$. The first list, $U(w)$, will be the list of all points in U that are on special lines through w (that is, lines containing w and at least two other distinct points). Notice that, since $w \notin \text{span}(U)$, each line through w can contain at most one point from U. The second list, $P_1(w)$, will be the list of points in $W \setminus \{w\}$ that are on a line containing w and a point from U. The third list, $P_2(w)$, will be of all other points on special lines through w (that is, on special lines that do not intersect U). These three lists are indeed disjoints, since w is the only common point between two lines passing through it. By the above discussion we have that $|P_1(w)| + |P_2(w)| \geq \delta m/2$ for all $w \in W$ (since removing these two lists destroys all collinear triples with w). We now separate the proof into two cases:

Case I : There exists $w \in W$ with $|P_1(w)| > \delta m/4$. In this case we can simply take U' to be the points in V that are also in the span of $\{w\} \cup U$. This new U' will include all points in $P_1(w)$ and so will grow by at least $\delta m/4$ points. Its dimension will grow by at most one and so we are done.

Case II : For all $w \in W$, $|P_2(w)| \geq \delta m/4$. Denote $m' = |W|$. In this case W itself is a δ'-LCC configuration with

$$\delta' = \frac{\delta m}{8m'}.$$

Applying Theorem 7.7 we get a sublist $U'' \subset W$ of size at least

$$\frac{(\delta')^3 m'}{K_1} \geq (\delta/8)^3 \cdot \frac{m}{K_1}$$

and dimension at most

$$\frac{K_1}{(\delta')^6} \leq K_1(8/\delta)^6.$$

We can thus take U' to be the points in V that are in the span of $U \cup U''$ and the proposition is proved. □

Proof of Theorem 7.6. We apply Proposition 7.11 on V, starting with the partition $U = \emptyset, W = V$ and ending when $U = V, W = \emptyset$. We can apply the proposition at most K_2/δ^3 times and in each step add at most K_2/δ^6 to the dimension of A (which is initially zero). Therefore, the final list $U = V$ will have dimension at most $O(1/\delta^9)$. □

7.2 Proof of Proposition 7.10

Order the points in V so that all copies of the same point are consecutive and so that $M(v_i) \leq M(v_j)$ whenever $i \leq j$. Let $S \subset V$ be the sublist containing the first $\delta m/10$ points in this ordering (we may be splitting the copies of a single point in the middle but this is fine). We will use the following simple fact later on:

Fact 7.12. *If $v \in S$ and $M(v') < M(v)$ then $v' \in S$.*

For a point $v \in V$ we denote by $T(v)$ the set of (ordered) triples in T containing v and for a line ℓ by $T_\ell(v)$ the set of (ordered) triples in T_ℓ containing v. Recall that these are all random variables determined by the choice of the mappings ρ for each line ℓ.

The proposition will follow by the following lemma.

Lemma 7.13. *Let $v \in S$. Then the expectation of $|T(v)|$ is at least $(\delta/10)^2 m$.*

The lemma completes the proof of the proposition: summing over all points in S we get

$$\begin{aligned}
\mathbb{E}[|T|] &\geq \mathbb{E}\left[(1/3)\sum_{v\in V}|T(v)|\right] \quad \text{(each triple is counted at most three times)} \\
&\geq (1/3)\sum_{v\in S}\mathbb{E}[|T(v)|] \\
&\geq (1/3)\cdot(\delta m/10)\cdot((\delta/10)^2 m) \geq (\delta/15)^3 m^2.
\end{aligned}$$

Proof of Lemma 7.13. Denote by $L(v)$ the set of all special lines through v. To prove the lemma we will identify a subfamily $L'(v)$ of $L(v)$ that contributes many triples to $T(v)$. To do so, we need the following definitions. For a set $\gamma \subset \mathbb{C}^d$ denote by $P(\gamma)$ the set of distinct points in V that are in γ. Denote $M(\gamma) = \sum_{v\in P(\gamma)} M(v)$. Denote by $P(\bar{S})$ the set of distinct points not in S.

Definition 7.14 (Degenerate line)**.** Let $\ell \in L(v)$. We say that $\ell \in L(v)$ is *degenerate* if either

1. The size of $P(\ell) \cap P(\bar{S})$ is at most one. That is, ℓ contains at most one distinct point outside S. Or,
2. There exists a point $v_\ell \in P(\ell)$, distinct from v, such that $M(v_\ell) \geq (1 - \delta/10)M(\ell)$.

A degenerate line satisfying the first (second) property above will be called a degenerate line of the first (second) kind.

Define $L'(v)$ as the set of line ℓ in $L(v)$ that are not degenerate. We will continue by proving two claims. The first claim shows that every line in $L'(v)$ contributes many triples in expectation to $T(v)$.

Claim 7.15. *For every $\ell \in L'(v)$ we have $\mathbb{E}[|T_\ell(v)|] \geq \delta M(\ell)/10$.*

Proof. Denote $r = |M(\ell)|$. The family of triples T_ℓ is obtained by taking a family of $r(r-1)$ triples F on $[r]$ (obtained from Lemma 4.6) and mapping it randomly to ℓ, omitting all triples that are not good (those that do not have three distinct points). For each triple $t \in F$ the probability that $\rho(t)$ will be in $T_\ell(v)$ can be lower bounded by

$$\frac{3}{r} \cdot \frac{2r}{3(r-1)} \cdot \frac{\delta}{20} = \frac{\delta}{10(r-1)}$$

The factor of $3/r$ comes from the probability that one of the three entries in t maps to v (these are disjoint events so we can sum their probabilities).

The next factor, $2r/(3(r-1))$, comes from the probability that the second entry in t (in some fixed order) maps to a point distinct from v. Indeed since $|P(\ell) \cap P(\bar{S})| \geq 2$ and using Fact 7.12 we know that there are at least two distinct points v', v'' on ℓ with $M(v') \geq M(v)$ and $M(v'') \geq v$. Since $M(v) + M(v') + M(v'') \leq r$, we get that $M(v) \leq r/3$, and so there are at least $2r/3$ "good" places for the second point to map to.

The last factor, $\delta/20$, comes from the probability that the third element of the triple will map to a point distinct from the first two. The bound of $\delta/20$ will follow from the fact that ℓ does not satisfy the second property in the definition of a degenerate line. To see why, let v_2 be the image of the second entry in t. Since ℓ is not degenerate, $r' \triangleq r - M(v_2) > \delta r/10$. Since $|P(\ell) \cap P(\bar{S})| \geq 2$, there is a point v' in $P(\bar{S})$ not in $\{v, v_2\}$, and hence, by Fact 7.12, $M(v) \leq M(v')$. Since $M(v) + M(v') \leq r'$, we get that $M(v) \leq r'/2$. Thus $r' - M(v) \geq r'/2 \geq \delta r/20$. But $r' - M(v)$ is exactly the number of 'good' places that the third entry can map to that are from v and v_2.

Using linearity of expectation we can conclude

$$\mathbb{E}[|T_\ell(v)|] \geq r(r-1) \cdot \frac{\delta}{10(r-1)} = \delta r/10.$$

□

The second claim shows that there are many points on lines in $L'(v)$.

Claim 7.16. *With the above notations, we have:*

$$\sum_{\ell \in L'(v)} M(\ell) \geq \delta m/10.$$

Proof. Assume in contradiction that

$$\sum_{\ell \in L'(v)} M(\ell) < \delta m/10.$$

Let Δ' denote the sub-list of V containing all points that lie on lines in $L'(v)$ so that $|\Delta'| \leq \delta m/10$. We will derive a contradiction by finding a small sublist Δ of V (containing Δ' and two other small sub-lists) that would violate Fact 7.8. That is, if we remove Δ from V, we destroy all collinear triples containing v.

Let ℓ be a degenerate line of the second kind. Then there is a point v_ℓ on it that is distinct from v and has multiplicity at least $(1-\delta/10)M(\ell)$. For every such line let Δ_ℓ denote the sublist of V containing all of the at most $(\delta/10)M(\ell) - M(v)$ points on this line that are distinct from both v and v_ℓ. Let Δ_2 denote the union of these lists Δ_ℓ over all degenerate lines of the second kind. We now have that $|\Delta_2| \leq \delta m/10$ since $\sum_\ell (M(\ell) - M(v)) \leq m$ and in each line ℓ we have

$$|\Delta_\ell| \leq (\delta/10)M(\ell) - M(v) \leq (\delta/10)(M(\ell) - M(v)).$$

Notice that, removing the points in Δ_2 destroys all collinear triples on degenerate lines of the second kind.

Finally, let Δ_S denote the sublist of V containing all points that have a copy in S. Thus Δ_S contains the list S (of at most $\delta m/10$ elements), plus all of the at most $\delta m/10$ copies of the last point in S, meaning that $|\Delta_S| \leq \delta m/5$. Removing Δ_S destroys all collinear triples on degenerate lines of the first kind. Define Δ as the union of the three sublists Δ', Δ_2 and Δ_S. From the above we have that removing Δ from V destroys all collinear triples containing V and that $|\Delta| \leq 4(\delta/10)m < \delta m/2$. This contradicts Fact 7.8. □

Combining the two claims we get that for all $v \in S$,

$$\mathbb{E}[|T(v)|] \geq \sum_{\ell \in L'(v)} \mathbb{E}[|T_\ell(v)|] \geq \sum_{\ell \in L'(v)} \delta M(\ell)/10 \geq (\delta/10) \cdot (\delta m/10) = (\delta/10)^2 m.$$

This completes the proof of Lemma 7.13. □

8 Extensions to Other Fields

In this section we show that our results can be extended from the complex field to fields of characteristic zero, and even to fields with very large positive characteristic. The argument is quite generic and relies on Hilbert's Nullstellensatz.

Definition 8.1 (T-matrix)**.** Let m, n be integers and let $T \subset [m] \times [n]$. We call an $m \times n$ matrix A a *T-matrix* if all entries of A with indices in T are non-zero and all entries with indices outside T are zero.

Theorem 8.2 (Effective Hilbert's Nullstellensatz [Kol88])**.** *Let $g_1, \ldots, g_s \in \mathbb{Z}[y_1, \ldots, y_t]$ be degree d polynomials with coefficients in $\{0, 1\}$ and let*

$$Z \triangleq \{y \in \mathbb{C}^t \mid g_i(y) = 0 \ \forall i \in [s]\}.$$

Suppose $h \in \mathbb{Z}[z_1, \ldots, z_t]$ is another polynomial with coefficients in $\{0, 1\}$ which vanishes on Z. Then there exist positive integers p, q and polynomials $f_1, \ldots, f_s \in \mathbb{Z}[y_1, \ldots, y_t]$ such that

$$\sum_{i=1}^{s} f_i \cdot g_i \equiv p \cdot h^q.$$

Furthermore, one can bound p and the maximal absolute value of the coefficients of the f_i's by an explicit function $H_0(d,t,s)$.

Theorem 8.3. *Let m,n,r be integers and let $T \subset [m] \times [n]$. Suppose that all complex T-matrices have rank at least r. Let $\mathbb{F}$ be a field of either characteristic zero or of finite large enough characteristic $p > P_0(n,m)$, where P_0 is some explicit function of n and m. Then, the rank of all T-matrices over $\mathbb{F}$ is at least r.*

Proof. Let $g_1, \ldots, g_s \in \mathbb{C}[\{x_{ij} \mid i \in [m], j \in [n]\}]$ be the determinants of all $r \times r$ sub-matrices of an $m \times n$ matrix of variables $X = (x_{ij})$. The statement "all T-matrices have rank at least r" can be phrased as "if $x_{ij} = 0$ for all $(i,j) \notin T$ and $g_k(X) = 0$ for all $k \in [s]$ then $\prod_{(i,j)\in T} x_{ij} = 0$." That is, if all entries outside T are zero and X has rank smaller than r then it must have at least one zero entry also inside T. From Nullstellensatz we know that there are integers $\alpha, \lambda > 0$ and polynomials $f_1, \ldots, f_s$ and $h_{ij}, (i,j) \notin T$, with integer coefficients such that

$$\alpha \cdot \left(\prod_{(i,j)\in T} x_{ij} \right)^{\lambda} \equiv \sum_{(i,j)\notin T} x_{ij} \cdot h_{ij}(X) + \sum_{k=1}^{s} f_i(X) \cdot g_i(X). \tag{2}$$

This identity implies the high rank of T-matrices also over any field $\mathbb{F}$ in which $\alpha \neq 0$. Since we have a bound on α in terms of n and m the result follows. □

9 Discussion and Open Problems

Our rank bound for design matrices has a dependence on q, the number of non-zeros in each row. Can this dependency be removed? This might be possible since a bound on q follows indirectly from specifying the bound on t, the sizes of the intersections. Removing this dependency might also enable us to argue about square matrices. Our results so far are interesting only in the range of parameters where the number of rows is much larger than the number of columns.

With respect to Sylvester-Gallai configurations, the most obvious open problem (discussed in the introduction) is to close the gap between our bound of $O(1/\delta^2)$ on the dimension of δ-SG configuration and the trivial lower bound of $\Omega(1/\delta)$ obtained by a simple partition of the points into $1/\delta$ lines.

Another interesting direction is to explore further the connection between design-matrices and LCCs. The most natural way to construct an LCC is by starting with a low-rank design matrix and then defining the code by taking the matrix to be its parity-check matrix. Call such codes *design-LCCs*. Our result on the rank of design matrices shows, essentially, that design-LCCs over the complex numbers cannot have good parameters in general (even for large query complexity). It is natural to ask whether there could exist LCCs that do not originate from designs. Or, more specifically, whether any LCC defines another LCC (with similar parameters) which is a design-LCC. This question was already raised in [BIW07]. Answering this question over the complex numbers will, using our results, give bounds for general LCCs. It is not out of the question to hope for bounds on LCCs with query complexity as large as polynomial in m (the encoding length). This would be enough to derive new results on rigidity via the connection

made in [Dvi10]. In particular, our results on design matrices still give meaningful bounds (on design-LCCs) in this range of parameters.

More formally, our results suggest a bound of roughly $\text{poly}(q, 1/\delta)$ on the dimension of (q, δ)-LCCs that arise from designs. A strong from of a conjecture from [Dvi10] says that an LCC $C \subset \mathbb{F}^n$ with $q = n^\epsilon$ queries and error $\delta = n^{-\epsilon}$, for some constant $\epsilon > 0$, cannot have dimension $0.99 \cdot n$. This conjecture, if true, would lead to new results on rigidity. Thus, showing that any LCC defines a design (up to some polynomial loss of parameters), combined with our results, would lead to new results on rigidity.

Acknowledgements

We thank Moritz Hardt for many helpful conversations. We thank Jozsef Solymosi for helpful comments.

References

[Alo09] Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Comb. Probab. Comput.*, 18(1-2):3–15, 2009.

[Bar98] Franck Barthe. On a reverse form of the brascamp-lieb inequality. *Inventiones Mathematicae*, 134:335–361, 1998. 10.1007/s002220050267.

[BE67] W. Bonnice and M. Edelstein. Flats associated with finite sets in $\mathbb{P}^d$. *Niew. Arch. Wisk.*, 15:11–14, 1967.

[BIW06] Boaz Barak, Russell Impagliazzo, and Avi Wigderson. Extracting randomness using few independent sources. *SIAM J. Comput.*, 36(4):1095–1118, 2006.

[BIW07] Omer Barkol, Yuval Ishai, and Enav Weinreb. On locally decodable codes, self-correctable codes, and t-private pir. In *APPROX '07/RANDOM '07: Proceedings of the 10th International Workshop on Approximation and the 11th International Workshop on Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 311–325, Berlin, Heidelberg, 2007. Springer-Verlag.

[BKT04] J. Bourgain, N. Katz, and T. Tao. A sum-product estimate in finite fields, and applications. *Geometric and Functional Analysis*, 14:27–57, 2004.

[BM90] P. Borwein and W. O. J. Moser. A survey of sylvester's problem and its generalizations. *Aequationes Mathematicae*, 40(1), 1990.

[CPR00] Bruno Codenotti, Pavel Pudlk, and Giovanni Resta. Some structural properties of low-rank matrices related to computational complexity. *Theoretical Computer Science*, 235(1):89 – 107, 2000.

[Dvi10] Zeev Dvir. On matrix rigidity and locally self-correctable codes. In *IEEE Conference on Computational Complexity*, pages 291–298, 2010.

[Erd43] P. Erdos. Problems for solution: 40654069, 1943.

[ES06] Lou M. Pretorius Elkies, Noam D. and Konrad J. Swanepoel. Sylvester-gallai theorems for complex numbers and quaternions,. *Discrete and Computational Geometry*, 35(3):361–373, 2006.

[FH07] Shaun M. Fallat and Leslie Hogben. The minimum rank of symmetric matrices described by a graph: A survey. *Linear Algebra and its Applications*, 426(2-3):558 – 582, 2007.

[For02] Jürgen Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.*, 65(4):612–625, 2002.

[Ham73] N. Hamada. On the p-rank of the incidence matrix of a balanced or partially balanced incomplete block design and its application to error correcting codes. *Hiroshima Math. J.*, 3:154–226, 1973.

[Han65] S. Hansen. A generalization of a theorem of sylvester on the lines determined by a finite point set. *Mathematica Scandinavia*, 16:175–180, 1965.

[Har10] Moritz Hardt. An algorithmic proof of forster's lower bound. Manuscript., 2010.

[Hil73] A. J. W. Hilton. On double diagonal and cross latin squares. *J. London Math. Soc.*, s2-6(4):679–689, 1973.

[JT09] Dieter Jungnickel and Vladimir D. Tonchev. Polarities, quasi-symmetric designs, and hamada's conjecture. *Des. Codes Cryptography*, 51(2):131–140, 2009.

[Kel86] L. M. Kelly. A resolution of the sylvester - gallai problem of j. -p. serre. *Discrete & Computational Geometry*, 1:101–104, 1986.

[Kol88] J. Kollr. Sharp effective nullstellensatz. *J. Amer. Math. Soc.*, 1:963–975, 1988.

[KS09] Neeraj Kayal and Shubhangi Saraf. Blackbox polynomial identity testing for depth 3 circuits. In *FOCS '09: Proceedings of the 2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 198–207, Washington, DC, USA, 2009. IEEE Computer Society.

[Lok09] Satyanarayana V. Lokam. Complexity lower bounds using linear algebra. *Foundations and Trends in Theoretical Computer Science*, 4(1-2):1–155, 2009.

[LSW00] N. Linial, A. Samorodnitsky, and A. Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. *Combinatorica*, 20(4):545–568, 2000.

[Mel40] E. Melchior. Uber vielseite der projektive ebene. *Deutsche Math.*, 5:461–475, 1940.

[RS89] U. Rothblum and H. Schneider. Scaling of matrices which have prespecified row sums and column sums via optimization. *Linear Algebra Appl*, 114-115:737–764, 1989.

[RS08] Alexander A. Razborov and Alexander A. Sherstov. The sign-rank of ac^o. In *FOCS '08: Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 57–66, Washington, DC, USA, 2008. IEEE Computer Society.

[Sin64] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

[SS10] Nitin Saxena and C. Seshadhri. From sylvester-gallai configurations to rank bounds: Improved black-box identity test for depth-3 circuits. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:21–29, 2010.

[ST83] Endre Szemerédi and William T. Trotter. Extremal problems in discrete geometry. *Combinatorica*, 3(3):381–392, 1983.

[Syl93] J. J. Sylvester. Mathematical question 11851. *Educational Times*, 59:98, 1893.

[Tre04] Luca Trevisan. Some applications of coding theory in computational complexity. *Quaderni di Matematica*, 13:347–424, 2004.

[Val77] Leslie G. Valiant. Graph-theoretic arguments in low-level complexity. In *MFCS*, pages 162–176, 1977.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de