Electronic Colloquium on Computational Complexity, Revision 2 of Report No. 34 (2012)

New Lower Bounds for Matching Vector Codes

Abhishek Bhowmick[*] Zeev Dvir[†] Shachar Lovett [‡]

Abstract

We prove new lower bounds on the encoding length of Matching Vector (MV) codes. These recently discovered families of Locally Decodable Codes (LDCs) originate in the works of Yekhanin [Yek08] and Efremenko [Efr09] and are the only known families of LDCs with a constant number of queries and sub-exponential encoding length. The systematic study of these codes, and their limitations, was initiated in [DGY11] where quasi-linear lower bounds were proved on their encoding length. Our work makes another step in this direction by proving two new lower bounds. The first is an unconditional quadratic lower bound, conjectured in [DGY11], which is the first bound to exceed the known lower bounds for general constant-query LDCs (when the number of queries is greater than four). The second result is a *conditional* super-polynomial lower bound for constant-query MV codes, assuming a well-known conjecture in additive combinatorics – the Polynomial Freiman Rusza conjecture (over $\mathbb{Z}_m^n$).

At the heart of MV codes are families of vectors in $\mathbb{Z}_m^n$ with restricted inner products modulo m. More precisely, families $U = (u_1, \ldots, u_t)$, $V = (v_1, \ldots, v_t)$ with $u_i, v_j \in \mathbb{Z}_m^n$ such that $\langle u_i, v_i \rangle = 0 \mod m$ for all $i \in [t]$ and $\langle u_i, u_j \rangle \neq 0 \mod m$ for $i \neq j$. Our lower bounds for MV codes are obtained by improving the known *upper bounds* on such families – a question that arises independently in combinatorics in the context of set systems with restricted modular intersections. In the course of our proofs we develop certain tools for working with matrices over $\mathbb{Z}_m$ that might be of independent interest.

[*] Department of Computer Science. University of Texas at Austin. Email: bhowmick@cs.utexas.edu. Research done while the author was a visiting student at Princeton University. Research supported by NSF grant CCF-0916160.

[†] Department of Computer Science and Department of Mathematics, Princeton University, Princeton NJ. 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, Princeton, NJ. Email: slovett@math.ias.edu. Research supported by NSF grant DMS-0835373.

ISSN 1433-8092

1 Introduction

Locally Decodable Codes (LDCs) are a special kind of Error Correcting Codes (ECCs) that allow the receiver to decode a *single* symbol of the message by querying a small number of positions in a corrupted encoding. More formally, an (r, δ, ϵ)-LDC encodes K-symbol messages x to N-symbol codewords $C(x)$, such that for every $i \in [K]$, the symbol x_i can be recovered with probability $1 - \epsilon$, by a randomized decoding procedure that makes only r queries, even if the codeword $C(x)$ is corrupted in up to δN locations. Since the early 90's, LDC's have found exciting applications in various areas ranging from data transmission to complexity theory to cryptography/privacy. We refer the reader to [Tre04, Yek11] for more background.

A central research question, which is far from being solved, has to do with understanding the best possible 'stretch' of an LDC with a constant number of queries. That is, how large N has to be as a function of K for constant r and with constant δ, ε (these two last parameters are not our focus here and we will generally assume them to be small fixed constants). For $r = 1, 2$ this question is completely answered. There are no LDC's for $r = 1$ [KT00] and the best LDC's with $r = 2$ have exponential encoding length [GKST02, KdW04]. For $r > 2$ there are huge gaps in our understanding. Katz and Trevisan were the first to study this problem [KT00] and, today, the best general lower bounds on N are slightly super-linear bounds of the form $\tilde{\Omega}\left(K^{1+1/(\lceil r/2 \rceil - 1)}\right)$ [Woo07]. Notice that, when the number of queries is 3 or 4, these bounds are quadratic (see also [KdW04, Woo10] for the $r = 3, 4$ case). The upper bounds were, until recently, those coming from polynomial codes and were of the order of $N \leq \exp\left(K^{\frac{1}{r-1}}\right)$. Improved upper bounds, breaking this barrier slightly, were given in [BIKR02].

This state of affairs changed dramatically when, in a breakthrough paper, Yekhanin [Yek08] developed a new approach for constructing LDCs that have much shorter codeword length than polynomial codes. Efremenko [Efr09] was the first to show that this approach could yield codes with subexponential encoding length (Yekhanin's paper showed this under a number theoretic assumption). More refinements and improvements to this new framework were obtained [Rag07, KY09, IS10, MFL$^+$10, DGY11, BET10] to give LDC's with r queries and with encoding length that grows, when r is a constant, roughly like

$$N \sim \exp\exp\left((\log K)^{O(\log\log r/\log r)}(\log\log K)\right).$$

While significantly smaller than the length of polynomial codes, the codeword length of these new codes is still super polynomial in K. In this work we prove that, assuming a well known conjecture in additive combinatorics – the Polynomial Freimann Ruzsa (PFR) conjecture – MV codes with constant number of queries must have super polynomial encoding length. We also show an unconditional quadratic lower bound on the length of MV codes which was conjectured in [DGY11] and holds also for large values of r (for which it is almost tight). We now describe our results in more detail.

1.1 Matching Vector Families

At the heart of every MV code there is a combinatorial object called a *Matching Vector family* (MV family). This is a pair or ordered lists $U = (u_1, \ldots, u_t), V = (v_1, \ldots, v_t)$, with each $u_i, v_i \in \mathbb{Z}_m^n$, where m, n are integers greater than 1. What makes U and V into an MV family are special properties of their inner products modulo m. These vectors must satisfy $\langle u_i, v_i \rangle = 0 \mod m$ for all $i \in [t]$ and $\langle u_i, v_j \rangle \neq 0 \mod m$ for all $i \neq j \in [t]$. The *size* of the matching vector family is the number of vectors – t. An MV family U, V in $\mathbb{Z}_m^n$ of size t is used to construct an MV code with $r = O(m)$ queries sending messages of length $K = t$ to codewords of length $N = m^n$.

It is important to note here that Yekhanin's original framework allowed for more general constructions that could, in some cases, reduce the number of queries further. For example, the codes given in [Yek08] use MV families over $\mathbb{Z}_p$, with p a Mersenne prime of magnitude going to infinity[1] and uses special properties of these primes to reduce the number of queries from $O(p)$ (as claimed above) to only 3. Another special case where the number of queries was reduced was in [Efr09] where (using a computer search) a certain gadget was found reducing the number of queries in an MV code over $\mathbb{Z}_{511}$ from the 'standard' 511 to 3. In this work, we refer to the 'basic' construction of MV codes (as described in [DGY11]) that does not use additional ad-hoc gadgets to reduce the number of queries. Thus, to prove a lower bound on the encoding length of MV codes, we will prove *upper bounds* on the size of an arbitrary MV families in $\mathbb{Z}_m^n$ for all m, n.

Let $\mathbf{MV}(m, n)$ denote the largest t such that there exists an MV family of size t in $\mathbb{Z}_m^n$. The question of bounding $\mathbf{MV}(m, n)$ is closely related to the well-known combinatorial problem of set systems with restricted modular intersections [BF98, Sga99, Gro00, Gro02] (in this setting the vectors u_i, v_i are required to have entries that are either 0 or 1). The systematic study of this more general problem, in the context of MV codes, was initiated in [DGY11]. When $m = p$ is prime, it is known that $\mathbf{MV}(p, n) \leq \min\{4p^{n/2} + 2, 1 + \binom{n+p-2}{p-1}\}$ [DGY11]. For general m, the best upper bound known [DGY11] is $\mathbf{MV}(m, n) \leq m^{n-1+o_m(1)}$, with $o_m(1)$ denoting a function that goes to zero when m grows. This bound is very weak when m (which bounds the number of queries of the corresponding MV code) is fixed and n grows. In particular, the resulting lower bound on the length of MV codes is only slightly super-linear. It was conjectured in [DGY11] that an upper bound of $\sim m^{n/2}$ should hold for any m (not just prime). However, the proof method used in [DGY11] to prove the $4p^{n/2} + 2$ bound does not extend to non primes. Notice that an upper bound of $\sim m^{n/2}$ on the size of MV families will imply a quadratic lower bound on the encoding length of MV codes.

Theorem 1. *For all $m, n \geq 2$ we have*

$$\mathbf{MV}(m, n) \leq m^{n/2 + O(\log m)}.$$

Hence, Theorem 1 resolves the conjecture of [DGY11] for the range $m = 2^{o(n)}$ and implies, in this range, a quadratic lower bound on the encoding length of MV codes with $O(m)$ queries. When $m >> n$, our bound is quite close to the best known construction of MV families which gives $\mathbf{MV}(m, n) \geq \left(\frac{m+1}{n-2}\right)^{n/2-1}$ [YGK12].

Our second result assumes the PFR conjecture (discussed below) and gives a stronger upper bound on the size of MV families when m is a constant and n grows.

Theorem 2. *Assuming the PFR conjecture over $\mathbb{Z}_m^n$ (Conjecture 1) we have*

$$\mathbf{MV}(m, n) \leq \exp\left(c(m)\frac{n}{\log n}\right),$$

[1] These are conjectured to exist.

with $c(m)$ an explicit function of m. Consequently, every MV code with $r = O(1)$ queries sending K bit messages to N bit codewords satisfies $N > K^{\Omega(\log\log(K))}$.

The PFR conjecture: Suppose $A \subseteq \mathbb{Z}_m^n$. The form of the PFR conjecture we will be using says that, if $|A - A| \leq \lambda \cdot |A|$ then there exists a subgroup H of $\mathbb{Z}_m^n$ of size at most $|A|$ such that A can be covered by at most $\mathrm{poly}(\lambda)$ translates of H. Stated formally:

Conjecture 1 (PFR Conjecture in $\mathbb{Z}_m^n$)**.** *Suppose $A \subseteq \mathbb{Z}_m^n$ and $|A - A| \leq \lambda \cdot |A|$. Then one can find a subgroup H of size at most $|A|$ such that A can be covered by $\lambda' = \lambda^{c_m}$ many translates of H, where c_m depends only on m.*

We note that the PFR conjecture has already found several applications in computer science. Ben-Sasson and Zewi [BSZ11] used it to construct two-source extractors from affine extractors; and Ben-Sasson, Lovett and Zewi [BSLZ11] used it to bound the deterministic communication complexity of functions whose corresponding matrix has low rank. Our work provides another application for the PFR and demonstrates its wide-reaching applicability. We further note that a quasi-polynomial version of the PFR conjecture was recently proved by Sanders [San10] (see also the exposition in [Lov12]). Unfortunately, all the applications discussed above require the truly polynomial version of the conjecture, and so cannot apply to Sanders' result.

From a technical point of view, one of the ingredients in this work builds on the recent work of Ben-Sasson, Lovett and Zewi [BSLZ11] who used the PFR conjecture to show that matrices over $\mathbb{Z}_2$ with large bias (say, with many more ones than zeros) and small rank must contain a large monochromatic sub-matrix. An important ingredient in our proof is a generalization of their results from $\mathbb{Z}_2$ to $\mathbb{Z}_m$ for all m, not necessarily prime. We note however that this is just one ingredient in our overall proof.

1.2 Proof Overview

The proof of Theorem 1 relies on intuitions coming from the theory of two-source extractors [CG88], which are functions of two variables $F(X, Y)$ such that the output of F is distributed in a close-to-uniform fashion whenever the two inputs are drawn, independently, from two distributions of sufficiently high entropy. Since our proof does not use two-source extractors explicitly we do not define them formally and just use them to explain the high level idea behind the proof. It is a well known fact [CG88] that the inner product function $F(X, Y) = \langle X, Y\rangle$, say over $\mathbb{Z}_2^n \times \mathbb{Z}_2^n$ is a good two source extractor when the two inputs X and Y are both drawn uniformly from sets $S_X, S_Y \subseteq \mathbb{Z}_2^n$ of size larger $2^{n/2}$. This immediately suggests a connection to MV families, since, if we take $S_X = U$ and $S_Y = V$ for an MV family U, V in $\mathbb{Z}_2^n$, we would get a completely non-uniform output (it will be zero with exponentially small probability). This means that the size of U, V is bounded from above by approximately $2^{n/2}$.

If we try to use a similar argument over $\mathbb{Z}_m$ we run into trouble since the inner product function modulo m is *not* a good two source extractors for sources of size $m^{n/2}$. Take, for example, $S_X = S_Y = \{0, 2, 4\}^n \subseteq \mathbb{Z}_6^n$ and observe that $\langle X, Y\rangle$ is always divisible by 2 and so is far from being uniformly distributed over $\mathbb{Z}_6$. It is, however, possible to show that this example is, in some sense, the only example and that, in general, we can always find a certain number of elements of either S_X or S_Y that 'agree' modulo some factor of m. This observation suggests proving Theorem 1 by induction on the number of factors of m, which is the way we proceed.

The proof of Theorem 2 uses a slightly different view of MV families as matrices with certain zero/non-zero pattern and small rank. Specifically, for an MV family U, V of size t in $\mathbb{Z}_m^n$ consider the $t \times t$ matrix P whose (i,j)'th entry is $\langle u_i, v_j \rangle \mod m$. The definition of an MV family implies that P has zeros on the diagonal and non-zeros everywhere else. If m was a prime, we could think of $\mathbb{Z}_m$ as a field $\mathbb{F}$ and say that, since P is the inner product matrix of vectors of length n over a field, it must have rank at most n. Conversely, every $t \times t$ matrix over a field $\mathbb{F}$ with these properties (zero on the diagonal and non-zero off the diagonal) and with rank n gives an MV family of size t in $\mathbb{F}^n$. We can call a matrix with this pattern of zeros/non-zeros an *MV matrix*. Thus, when m is prime, the question of bounding the size of an MV family is the same as lower bounding the rank of an MV-matrix[2]. When m is composite, this whole approach should be re-examined since $\mathbb{Z}_m$ is no longer a field and our familiar understanding of matrices and linear algebra over a field are no longer valid. We do, however, manage to carry over this correspondence between the two problems by defining the notion of rank in a careful way (more on this issue below).

Assume for the purpose of this overview that the usual notion of rank and other intuitions from linear algebra are valid over $\mathbb{Z}_m$ and let us proceed with sketching the proof of Theorem 2 using the equivalent formulation as bounding (from below) the rank of an MV matrix P. The starting point is a generalization of a result of [BSLZ11], mentioned above, from $\mathbb{Z}_2$ to $\mathbb{Z}_m$. We show that every matrix P over $\mathbb{Z}_m$ that is biased (i.e., its values are not distributed close to uniformly) and has low rank, contains a large monochromatic sub-matrix *modulo some factor m' of m*. The size of the sub-matrix is bounded from below by $\sim |P| \exp(r'/\log(r'))$, where r' is the rank of P modulo m' (this factor depends on the specific way the matrix is biased). This generalizes the result of [BSLZ11] which assumes $m = 2$ and finds a large monochromatic sub-matrix (modulo 2). Let us refer to this result from now on as the *sub-matrix* lemma. We can apply the sub-matrix lemma to an MV matrix P since its values are far from uniform (the probability of zero is much less than $1/m$) and since its rank is assumed (towards a contradiction) to be low.

Suppose for the sake of simplicity that $m = p \cdot q$, with p, q distinct primes (the proof for general m is significantly more technical but relies on the same basic intuitions). Applying the sub-matrix lemma we obtain a sub-matrix P_1 of P that is constant modulo some factor m_1 of m (so m_1 is either p, q or m) of size at least $|P| \exp(-r_1/\log(r_1))$, where $r_1 \leq n$ is the rank of $P \mod m_1$. Using some matrix manipulations, and subtracting a rank one matrix, we can get a large sub-matrix P_1' that does not intersect the diagonal of P and s.t all of the entries of P_1' are zero modulo m_1. Suppose $|P_1'| = t_1$ and consider the $2t_1 \times 2t_1$ sub-matrix P_1'' of P that has P_1' as its top-right (or bottom-left) block and s.t the top-left and bottom-right blocks are taken to have zero diagonal elements. Formally, if P_1' is indexed by rows in R and columns in T with $R \cap T = \emptyset$ then the rows/columns of P_1'' will be indexed by $R \cup T$. If we consider the matrix P_1'' modulo m_1 then it has top-right block which is all zero and so its rank (modulo m_1) will be the sum of the ranks of the top-left and bottom right blocks. Thus, one of these blocks, w.l.o.g the top-left one, must have rank at most $n/2$ (over $\mathbb{Z}_{m_1}$). Notice also that both of these blocks are themselves MV matrices modulo m since they are sub-matrices of P with the same row and column sets. Let $\tilde{P}_1$ be the top-left block of P_1''. We can now apply, again, the monochromatic sub-matrix lemma to find a

[2] For technical reasons, the actual proof will not be entirely using matrices and will keep the MV family in the background. This is because we need to keep certain invariants throughout the proof and these are easier to define for families of vectors than for matrices.

large sub-matrix P_2 of $\tilde{P}_1$ which is constant modulo some other factor m_2 of m. The size of P_2 will be

$$t_1 \cdot \exp(-r_2/\log(r_2)) = |P| \cdot \exp(-r_1/\log(r_1) - r_2/\log(r_2)).$$

The factor m_2 is also either p or q. If it happens to be that $m_1 = m_2$ then $r_2 \leq n/2$ and so we gain in the size of P_2 in this second step (the expression $r_2/\log(r_2)$ is smaller than $n/2\log(n/2)$ which is smaller by roughly a factor of two than our bound on $r_1/\log(r_1)$. Suppose we continue with this iterative process of finding constant sub-matrices for ℓ steps and that, by luck, all the factors $m_1, m_2, \ldots$ are equal to the same factor of m (say p). Then, after roughly $\log(n)$ iteration, we will reduce the rank modulo p to one and still have at least

$$|P| \cdot \exp\left(-\sum_{i=1}^{\ell} \frac{n}{2^i \log(n/2^i)}\right)$$

rows, which is close to the original size of P if we assume (in contradiction) that $|P| >> \exp(n/\log n)$ (this calculation is given in Claim A.1). In this case we obtain a new large MV family U', V' modulo m such that all inner products $\langle u'_i, v'_j \rangle$ of elements $u'_i \in U', v'_j \in V'$ are fixed modulo p. From this we can easily construct an MV family of roughly the same size in $\mathbb{Z}_q^n$ and then use the bounds on $\mathbf{MV}(q, n)$ for primes to get a contradiction. In the 'unlucky' case we will have different factors $m_1, m_2, \ldots$ in each stage, but we can adapt the analysis to consider the decrease in rank simultaneously for all factors of m.

The full proof is by induction on the number of factors of m and uses the iterative sub-matrix argument above to go from an MV family modulo m to an MV family of roughly the same size modulo some proper factor of m (and then uses the inductive hypothesis on this new MV family).

1.3 Matrix rank over $\mathbb{Z}_m$

An important technical issue, which was already hinted at above, is in the definition of the rank of a matrix with entries in a ring $\mathbb{Z}_m$. There are two main properties of matrix rank over a field that we relied on in the proof sketch above. The first is that a rank r matrix is always the inner product matrix of vectors in r dimensions. Equivalently, a $t \times t$ matrix of rank r can be written as a product of a $t \times r$ matrix and an $r \times t$ matrix. This is important if we are to go back and forth between matrices and MV families. Another property we used is that, if we have a $2t \times 2t$ matrix composed of 4 blocks of size $t \times t$ and the top-right block is zero, then the rank of the matrix is the sum of the ranks of the top-left block and the bottom right block.

Ideally, we would like to define rank over $\mathbb{Z}_m$ so that both properties are satisfied. This is, however, impossible as the following example shows: Consider the 2×2 matrix with the two rows $(4, 0)$ and $(0, 3)$ over $\mathbb{Z}_6$. This matrix can be written as the product of the two vectors $(2, 3)^T$ and $(2, 3)$ and so should have rank one, if we are to satisfy the first property. However, if we are to satisfy the second property, its rank should be the sum of the ranks of the two 1×1 matrices (4) and (3), which clearly cannot have rank zero!

Our solution to this problem is to give two different definitions of rank, each satisfying one of the two properties. We then show that the two definitions of rank can differ from each other by a multiplicative factor of $\log m$, which our proof can handle. The first definition of rank is as the smallest r such that our $t \times t$ matrix can be written as a product of a $t \times r$ matrix and an $r \times t$

matrix. Clearly this would satisfy the first property (but not the second). The second definition of rank is termed *column-rank* and is defined as the logarithm to the base m of the size of the additive subgroup of $\mathbb{Z}_m^t$ generated by the columns of the matrix. Notice that this definition of rank can result in the rank being non-integer. For example, the rank of the matrix with a single column $(2, 0)$ over $\mathbb{Z}_6$ would be equal to $\log_6(3)$ since the subgroup generated by this column is composed of the three vectors $(2, 0), (4, 0), (0, 0)$. It is not hard to show (see Claim 4.9) that this definition satisfies the second property described above regarding block matrices. Clearly, the two definitions agree for matrices over a field. We show (see Claim 4.6) that the two notions of rank can differ by a multiplicative factor of at most $\log m$. This allows us to use both definitions in different parts of the proof without losing too much in the transition. We finish this discussion by noting that in no part of the proof do we use the characterization of rank using determinants, which is often very useful when working over a field.

1.4 Organization

We begin with some preliminaries in Section 2. We prove Theorem 1 in Section 3. Section 4 contains some claims about matrices over $\mathbb{Z}_m$. Section 5 introduces collision free MV families. Both Section 4 and Section 5 will be used in the proof of Theorem 2 in Section 6. The proof of Theorem 2 also requires the sub-matrix lemma, whose proof appears in Section 7.

2 General preliminaries

Notations: Throughout the paper we will be handling ordered lists of elements. A list A of size t over a finite set Ω is an ordered t-tuple $A = (a_1, a_2, \cdots, a_t)$ where each $a_i \in \Omega$. A list can have repetitions. If it doesn't we say it is *twin free*. When discussing sublists $A \subseteq B$ with $B = (b_1, \ldots, b_t)$ we will use the convention that, unless specified otherwise, A maintains the ordering induced by B. For a positive integer t, we let $[t]$ denote the list $(1, \cdots t)$. So, for example, when we say that $T \subseteq [t]$ we mean that T is a list of integers in increasing order belonging to $[t]$. We say that a list $A = (a_1, \ldots, a_t)$ over Ω is *constant* if $a_i = a_j$ for all $i, j \in [t]$. We assume all logarithms are in base 2 unless otherwise specified.

2.1 MV Families: Basic Facts and Definitions

We now start with some basic definition and claims regarding MV families.

Definition 2.1 (Matching Vector Family). *Let $U = (u_1, u_2, \cdots u_t)$ and $V = (v_1, v_2, \cdots v_t)$ be lists over $\mathbb{Z}_m^n$. Then (U, V) is called a* matching vector family *of size t in $\mathbb{Z}_m^n$ if*

- $\langle u_i, v_i \rangle = 0 \ \ (mod\ m)\,, \quad \forall i.$
- $\langle u_i, v_j \rangle \neq 0 \ \ (mod\ m)\,, \quad \forall i \neq j.$

We denote the size of (U, V) by $|(U, V)|$. For instance, $|(U, V)| = t$ above.

Definition 2.2 (Subset of Matching Vector Family). *Let $U = (u_1, u_2, \cdots u_t)\,, V = (v_1, v_2, \cdots v_t)$ form a matching vector family in $\mathbb{Z}_m^n$ of size t. By $(U', V') \subseteq (U, V)$, we mean there exists a sublist*

$T \subseteq [t]$ such that $U' = (u_i : i \in T), V' = (v_i : i \in T)$. Observe that (U', V') is a matching vector family in $\mathbb{Z}_m^n$.

Definition 2.3 ($\mathbf{MV}(m, n)$)**.** *We denote by $\mathbf{MV}(m, n)$ the maximum size of a matching vector family (U, V) in $\mathbb{Z}_m^n$.*

We shall use the following simple facts implicitly throughout the paper.

Fact 2.4. *$\mathbf{MV}(m, n)$ is an increasing function of n.*

Proof. For $n_1 < n_2$, we show $\mathbf{MV}(m, n_1) \le \mathbf{MV}(m, n_2)$. Given (U, V), a matching vector family in $\mathbb{Z}_m^{n_1}$, we can pad each element in U and V by $n_2 - n_1$ zeros and obtain a matching vector family in $\mathbb{Z}_m^{n_2}$ of the same size. □

Fact 2.5. *If (U, V) is a matching vector family in $\mathbb{Z}_m^n$, then U and V are twin free.*

Proof. Let $U = (u_1, u_2, \cdots u_t), V = (v_1, v_2, \cdots v_t)$. We prove U is twin free. By symmetry V is also twin free. Suppose $u_i = u_j$ for some $i \neq j$. Now, $\langle u_i, v_j \rangle = \langle u_j, v_j \rangle = 0$ which is a contradiction. □

To facilitate writing in the proofs to follow we introduce the following notation for taking lists, matrices, etc. modulo an integer r.

Definition 2.6 (Modulo r notation)**.** *Let $2 \le r \le m$ be such that r divides m. Given $a = (a_1, \cdots, a_n) \in \mathbb{Z}_m^n$, we denote by $a^{(r)} = (a_1 \pmod r, \cdots, a_n \pmod r) \in \mathbb{Z}_r^n$. For a list $U = (u_1, u_2, \cdots u_t)$ over $\mathbb{Z}_m^n$, let $U^{(r)} = \left(u_1^{(r)}, u_2^{(r)}, \cdots u_t^{(r)}\right)$. Also, if $u^{(r)}$ is constant for all $u \in U$, we say $U^{(r)}$ is constant. Similarly, for a $t \times t$ matrix M over $\mathbb{Z}_m$, define $M^{(r)}$ to be the $t \times t$ matrix over $\mathbb{Z}_r$ such that $M^{(r)}(j, k) = M(j, k) \pmod r$ for all $1 \le j, k \le t$.*

We will also need the following definitions.

Definition 2.7 (Bucket $B_r(w, A)$)**.** *Let $A \subseteq \mathbb{Z}_m^n$ be a list. For any $w \in \mathbb{Z}_r^n$, we denote by $B_r(w, A) = \left(a \in A : a^{(r)} = w\right)$ the sub-list of elements of A which are equal to w modulo r.*

Definition 2.8 (Matrix $P_{U,V}$)**.** *Let $U = (u_1, u_2, \cdots u_t)$ and $V = (v_1, v_2, \cdots v_t)$ be lists over $\mathbb{Z}_m^n$. We let $P_{U,V}$ be the $t \times t$ matrix over $\mathbb{Z}_m$ defined by $P_{U,V}(i, j) = \langle u_i, v_j \rangle$ for $1 \le i, j \le t$.*

We will use the following lemma from [DGY11] mentioned informally in the introduction.

Lemma 2.9. *[DGY11, Theorem 21] For any positive integer n and prime p, $\mathbf{MV}(p, n) \le 1 + \binom{n+p-2}{p-1}$.*

2.2 Probability Distributions

Definition 2.10. *For a distribution μ over a finite set Ω, we write $X \sim \mu$ to denote a random variable X drawn according to μ. We will also treat μ as a function $\mu : \Omega \mapsto [0, 1]$ such that $\mu(x) = \mathbf{Pr}[X = x]$. For a list A over Ω, $x \sim A$ denotes a point sampled as per the uniform distribution on A (taking repetitions into account).*

Definition 2.11 (Statistical distance between distributions)**.** *Let μ_1 and μ_2 be two distributions over a finite set Ω. The* statistical distance *(or simply distance) between μ_1 and μ_2, denoted $\Delta(\mu_1, \mu_2)$, is defined as*

$$\Delta(\mu_1, \mu_2) = \frac{1}{2} \sum_{x \in \Omega} |\mu_1(x) - \mu_2(x)| .$$

Definition 2.12 (Collision probability)**.** *Given a distribution μ over a finite set Ω the* collision probability *of μ, denoted* $\mathrm{cp}(\mu)$, *is defined as*

$$\mathrm{cp}(\mu) = \mathbf{Pr}_{x,y \sim \mu}[x = y] = \sum_{x \in \Omega} \mu(x)^2 .$$

The following two lemmas are standard and their proofs are included, for completeness, in Appendix B.

Lemma 2.13. *Let μ be a distribution over $\mathbb{Z}_m$ and let $\mathcal{U}_m$ denote the uniform distriution over $\mathbb{Z}_m$. If $\Delta(\mu, \mathcal{U}_m) \geq \epsilon$ then for some $1 \leq j \leq m-1$,*

$$\left| \mathbb{E}_{x \sim \mu} \left[\left(\omega^j \right)^x \right] \right| \geq \frac{2\epsilon}{\sqrt{m}},$$

where $\omega = exp(2\pi i/m)$ is a primitive root of unity of order m.

Lemma 2.14. *Let ω be a primitive root of unity of order m. Let μ_1 and μ_2 be two probability distributions over $\mathbb{Z}_m^n$. If $\left| \mathbb{E}_{x \sim \mu_1, y \sim \mu_2} \left[\omega^{\langle x,y \rangle} \right] \right| \geq \epsilon$, then* $\mathrm{cp}(\mu_1)\,\mathrm{cp}(\mu_2) \geq \epsilon^2/m^n$.

3 Proof of Theorem 1

In this section we prove Theorem 1, restated here with explicit constants.

Theorem 3.1. *Let $m \geq 2$ and n be arbitrary positive integers. Then*

$$\mathbf{MV}(m, n) \leq m^{n/2 + 14 \log m}.$$

For the purpose of the proof, we introduce a notation that will be used only in this section.

Definition 3.2 ($\mathbf{MV}_{r_1,r_2}(m, n)$)**.** *Let r_1, r_2 be integers such that $r_1 r_2 | m$. We denote by $\mathbf{MV}_{r_1,r_2}(m, n)$ the maximum size of an MV family (U, V) in $\mathbb{Z}_m^n$ satisfying*

- *$U^{(r_1)}$ and $V^{(r_2)}$ are constants.*
- *$\langle u, v \rangle = 0 \pmod{r_1 r_2}$ for all $u \in U, v \in V$.*

Note that $\mathbf{MV}_{1,1}(m, n) = \mathbf{MV}(m, n)$ (with the convention that $x \pmod 1 = 0$ for any integer x).

The proof of Theorem 3.1 will follow immediately from the following two lemmas, which will be proved below.

Lemma 3.3. *Let $m = r_1 r_2 r_3$ where r_1, r_2, r_3 are arbitrary positive integers such that $r_3 \geq 2$. Let $t \geq 3m$ and $n \geq 1$ be arbitrary integers. Let (U, V) be a matching vector family in $\mathbb{Z}_m^n$ with $|(U, V)| = t$ such that*

- *$U^{(r_1)}$ and $V^{(r_2)}$ are constants.*
- *$\langle u, v\rangle = 0 \pmod{r_1 r_2}$ for all $u \in U, v \in V$.*

Then, there exists $s|r_3$ with $s \geq 2$ and a matching vector family $(U', V') \subseteq (U, V)$ such that $|(U', V')| \geq s^{-n/2} m^{-11} t$ where

- *$\langle u', v'\rangle = 0 \pmod{r_1 r_2 s}$ for all $u' \in U', v' \in V'$.*
- *Either $U'^{(r_1 s)}$ is constant or $V'^{(r_2 s)}$ is constant.*

Applying Lemma 3.3 iteratively we prove the following bound.

Lemma 3.4. $\mathbf{MV}_{r_1,r_2}(m, n) \leq 3m \cdot m^{11 \log \frac{m}{r_1 r_2}} \left(\frac{m}{r_1 r_2}\right)^{n/2}$.

Given Lemma 3.4 the proof of Theorem 3.1 is immediate.

Proof of Theorem 3.1. Observe that for any matching vector family (U, V) in $\mathbb{Z}_m^n$, $U^{(1)}$ and $V^{(1)}$ are constants and $\langle u, v\rangle = 0 \pmod{1}$ for all $u \in U, v \in V$. Thus, setting $r_1 = r_2 = 1$ in Lemma 3.4 implies $\mathbf{MV}(m, n) \leq 3m \cdot m^{11 \log m} m^{n/2} < m^{n/2 + 14 \log m}$. □

3.1 Proof of Lemma 3.3

By assumption we have that $\langle u, v\rangle = 0 \pmod{r_1 r_2}$ for all $u \in U, v \in V$. So, we can consider $\frac{\langle u,v\rangle}{r_1 r_2} \in \mathbb{Z}_{r_3}$. We have that

- For $1 \leq i \leq t$, $\frac{\langle u_i,v_i\rangle}{r_1 r_2} = 0 \pmod{r_3}$ since $\langle u_i, v_i\rangle = 0 \pmod{m}$.
- For $1 \leq i, j \leq t$, $i \neq j$, $\frac{\langle u_i,v_j\rangle}{r_1 r_2} \neq 0 \pmod{r_3}$ since $\langle u_i, v_j\rangle \neq 0 \pmod{m}$.

Let μ denote the distribution over $\mathbb{Z}_{r_3}$ defined by $\frac{\langle u_i,v_j\rangle}{r_1 r_2} \bmod r_3$ where u_i, v_j are drawn independently and uniformly from U, V respectively. Observe that μ outputs 0 only when $i = j$. Therefore, $\mathbf{Pr}[\mu = 0] = 1/t \leq 1/3m$. On the other hand, $\mathbf{Pr}[\mathcal{U}_{r_3} = 0] = 1/r_3 \geq 1/m$. This implies that $\Delta(\mu, \mathcal{U}_{r_3}) \geq 1/3m$. Thus, applying Lemma 2.13 with $\omega = exp(2\pi i/r_3)$, we get that for some $1 \leq j \leq r_3 - 1$,

$$\left|\mathbb{E}_{x\sim\mu}\left[\left(\omega^j\right)^x\right]\right| \geq \frac{2}{3m\sqrt{r_3}} \geq \frac{2}{3m^{3/2}}.$$

Let $\omega' = \omega^j$ and $s = r_3/gcd(r_3, j)$ be the order of ω'. Also, note that as $j \geq 1$, we have $s \geq 2$. Let $\varepsilon = \frac{2}{3m^{3/2}}$. Using the Cauchy-Schwartz inequality twice we get

$$\begin{aligned}
& \left|\mathbb{E}_{u\sim U, v\sim V}\left[\left(\omega'\right)^{\langle u,v\rangle/r_1 r_2}\right]\right| \geq \epsilon \\
\Longrightarrow\quad & \left|\mathbb{E}_{u,\tilde{u}\sim U, v\sim V}\left[\left(\omega'\right)^{\langle u-\tilde{u},v\rangle/r_1 r_2}\right]\right| \geq \epsilon^2 \\
\Longrightarrow\quad & \left|\mathbb{E}_{u,\tilde{u}\sim U, v,\tilde{v}\sim V}\left[\left(\omega'\right)^{\langle u-\tilde{u},v-\tilde{v}\rangle/r_1 r_2}\right]\right| \geq \epsilon^4 \\
\Longrightarrow\quad & \left|\mathbb{E}_{u,\tilde{u}\sim U, v,\tilde{v}\sim V}\left[\left(\omega'\right)^{\langle (u-\tilde{u})/r_1,(v-\tilde{v})/r_2\rangle}\right]\right| \geq \epsilon^4.
\end{aligned}$$

We need to explain the last expression. Since by assumption $U^{(r_1)}$ and $V^{(r_2)}$ are constants, $(u-\tilde{u})/r_1 \in \mathbb{Z}_m^n$ and $(v-\tilde{v})/r_2 \in \mathbb{Z}_m^n$ are well defined. Thus, we can fix $\tilde{u}$ and $\tilde{v}$ by an averaging argument such that

$$\left|\mathbb{E}_{u\sim U, v\sim V}\left[\left(\omega'\right)^{\langle (u-\tilde{u})/r_1, (v-\tilde{v})/r_2\rangle}\right]\right| \geq \epsilon^4.$$

Let $U' = (u'_1, u'_2, \cdots u'_t), V' = (v'_1, v'_2, \cdots v'_t)$ where $u'_i = (u_i - \tilde{u})/r_1$ and $v'_i = (v_i - \tilde{v})/r_2$. Notice that U' and V' are not assumed to be an MV family (later we will derive from them an MV family). We now define two probability distributions $\mu^{U'}$ and $\mu^{V'}$ over $\mathbb{Z}_s^n$. For each $w \in \mathbb{Z}_s^n$, let $\mu^{U'}(w) = |B_s(w, U')| / |U'|$ and $\mu^{V'}(w) = |B_s(w, V')| / |V'|$. That is, $\mu^{U'}(w)$ is the probability that $u'^{(s)} = w$ where u' is chosen uniformly in U', and similarly for $\mu^{V'}(w)$. Therefore, since the order of w' is s, we have that

$$\left|\mathbb{E}_{w_1\sim\mu^{U'}, w_2\sim\mu^{V'}}\left[\left(\omega'\right)^{\langle w_1, w_2\rangle}\right]\right| \geq \epsilon^4.$$

Recalling that s is the order of ω' and applying Lemma 2.14, we get $\mathrm{cp}\left(\mu^{U'}\right)\mathrm{cp}\left(\mu^{V'}\right) \geq \epsilon^8/s^n$. Therefore, one of $\mathrm{cp}\left(\mu^{U'}\right)$, $\mathrm{cp}\left(\mu^{V'}\right)$, say $\mathrm{cp}\left(\mu^{U'}\right)$, is at least $\epsilon^4/s^{n/2}$. Let w^* be the point of maximum probability mass given by $\mu^{U'}$. Then,

$$\mu^{U'}(w^*) = \mu^{U'}(w^*)\sum_{w\in\mathbb{Z}_s^n}\mu^{U'}(w) \geq \sum_{w\in\mathbb{Z}_s^n}\mu^{U'}(w)^2 = \mathrm{cp}\left(\mu^{U'}\right) \geq \epsilon^4/s^{n/2}.$$

Now, $\mu^{U'}(w^*) \geq \epsilon^4/s^{n/2}$ means that $\left|\left\{u \in U : \frac{u-\tilde{u}}{r_1} = w^* \;\; (mod\ s)\right\}\right| \geq t\epsilon^4/s^{n/2}$. Equivalently,

$$\left|\Big\{u \in U : u - \tilde{u} = r_1 w^* \;\; (mod\ r_1 s)\Big\}\right| \geq t\epsilon^4/s^{n/2}.$$

Let $T' = (i : u_i = \tilde{u} + r_1 w^* \;\; (mod\ r_1 s))$. Now, define $U'' = (u_i : i \in T')$ and $V'' = (v_i : i \in T')$. Observe that (U'', V'') is a matching vector family in $\mathbb{Z}_m^n$ such that

- $U''^{(r_1 s)}$ and $V''^{(r_2)}$ are constants.
- $|(U'', V'')| \geq t\left(\epsilon^4/s^{n/2}\right)$.

The only thing left is to show that $\langle u, v\rangle = 0 \;\; (mod\ r_1 r_2 s)$ for all $u \in U'', v \in V''$. This may not be true in general. However, we can take a large subset of the matching vector family so that the resulting matching vector family satisfies this condition. To see this, let $u \in U'', v \in V''$ be arbitrary. Now, $u = r_1 s \cdot u' + u_0$ and $v = r_2 \cdot v' + v_0$ where u', v' depend on u, v respectively and u_0, v_0 are independent of u, v. Then,

$$\langle u, v\rangle = r_1 r_2 s\langle u', v'\rangle + r_1 s\langle u', v_0\rangle + r_2\langle u_0, v'\rangle + \langle u_0, v_0\rangle.$$

As u varies over U'', $\langle u', v_0\rangle$ takes at most r_2 values modulo r_2. Hence, $r_1 s\langle u', v_0\rangle$ takes at most $r_2 \leq m$ values modulo $r_1 r_2 s$. Therefore, there exist at least $(1/m)|U''|$ elements of U'' such that $r_1 s\langle u', v_0\rangle$ is a constant modulo $r_1 r_2 s$. We take the corresponding elements from V'' to form a matching vector family $(U''', V''') \subseteq (U'', V'')$. We apply another round using the same idea on U''', V''', this time ensuring that $r_2\langle u_0, v'\rangle$ is constant modulo r_1r_2s as v varies over a large fraction of V'''. Thus, we end up with $\tilde{V}$ of size at least $(1/m)\,|V'''|$ such that $r_2\langle u_0, v_i\rangle$ is a constant modulo r_1r_2s. We take the corresponding subset $\tilde{U}$ from U''' so that $(\tilde{U}, \tilde{V}) \subseteq (U''', V''')$ is a matching vector family. Denote the size of $(\tilde{U}, \tilde{V})$ by $\tilde{t}$. Note that $\tilde{U} = (\tilde{u}_1, \cdots, \tilde{u}_{\tilde{t}})\,, \tilde{V} = (\tilde{v}_1, \cdots, \tilde{v}_{\tilde{t}})$ is a matching vector family in $\mathbb{Z}_m^n$ of size at least $\left(1/m^2\right) t \left(\epsilon^4/s^{n/2}\right) = s^{-n/2} m^{-(8+\log_m(81/16))} t \geq s^{-n/2} m^{-8-\log(81/16)} t \geq s^{-n/2} m^{-11} t$. Also, as $\langle u, v\rangle$ is a constant modulo r_1r_2s, for $u \in \tilde{U}, v \in \tilde{V}$, and $\langle \tilde{u}_i, \tilde{v}_i\rangle = 0 \pmod{r_1r_2s}$, we get that $\langle u, v\rangle = 0 \pmod{r_1r_2s}$, for $u \in \tilde{U}, v \in \tilde{V}$. This concludes the proof. □

3.2 Proof of Lemma 3.4

We prove the lemma by backward induction on $r_1r_2|m$. That is, to prove the claim about $\mathbf{MV}_{r_1,r_2}(m,n)$, we assume the inductive hypothesis for $\mathbf{MV}_{r_1',r_2'}(m,n)$ where $r_1'r_2' > r_1r_2$ and $r_1'r_2'|m$.

Base Case. The base case of $r_1r_2 = m$ is trivial. To see this, observe that if $\langle u, v\rangle = 0 \pmod{m}$ for all $u \in U, v \in V$, then by the definition of a matching vector family in $\mathbb{Z}_m^n$, the size of such a family cannot exceed 1. Hence, for $r_1r_2 = m$, $\mathbf{MV}_{r_1,r_2}(m,n) = 1 \leq 3m \cdot m^{c \log \frac{m}{r_1r_2}} \left(\frac{m}{r_1r_2}\right)^{n/2}$.

Inductive Step. Let $m = r_1r_2r_3$ with $r_1r_2 < m$ (that is, $r_3 \geq 2$). By the inductive hypothesis we have $\mathbf{MV}_{r_1',r_2'}(m,n) \leq 3m \cdot m^{c \log \frac{m}{r_1'r_2'}} \left(\frac{m}{r_1'r_2'}\right)^{n/2}$ for all r_1', r_2' such that $r_1'r_2' > r_1r_2$ and $r_1'r_2'|m$. We need to show that $\mathbf{MV}_{r_1,r_2}(m,n) \leq 3m \cdot m^{c \log \frac{m}{r_1r_2}} \left(\frac{m}{r_1r_2}\right)^{n/2}$. Suppose this is false, so that there exists a matching vector family (U, V) in $\mathbb{Z}_m^n$ with $U = (u_1, \cdots u_t)\,, V = (v_1, \cdots v_t)$ where $t > 3m \cdot m^{c \log \frac{m}{r_1r_2}} \left(\frac{m}{r_1r_2}\right)^{n/2}$ such that

- $U^{(r_1)}$ and $V^{(r_2)}$ are constants.
- $\langle u, v\rangle = 0 \pmod{r_1r_2}$ for all $u \in U, v \in V$.

Note that $t \geq 3m$. Therefore, applying Lemma 3.3, there exists $s|r_3$ with $s \geq 2$ and matching vector family $(U', V') \subseteq (U, V)$ such that $|(U', V')| \geq s^{-n/2} m^{-11} t$ where

- $\langle u', v'\rangle = 0 \pmod{r_1r_2s}$ for all $u' \in U', v' \in V'$.
- either $U'^{(r_1s)}$ is constant or $V'^{(r_2s)}$ is constant.

Without loss of generality, we assume that $U'^{(r_1s)}$ is a constant. Therefore,

$$\begin{aligned}\left|(U', V')\right| &> s^{-n/2} m^{-11} \cdot 3m \cdot m^{11 \log \frac{m}{r_1r_2}} \left(\frac{m}{r_1r_2}\right)^{n/2} \\ &= 3m \left(\frac{m}{r_1r_2s}\right)^{n/2} m^{11\left(\log \frac{m}{r_1r_2} - 1\right)} \\ &\geq 3m \left(\frac{m}{r_1r_2s}\right)^{n/2} m^{11 \log \frac{m}{r_1r_2s}},\end{aligned}$$

where the last inequality used the fact that $s \geq 2$. This however contradicts the inductive hypothesis. $\square$

4 Matrices over $\mathbb{Z}_m$

Notations: For a $t \times s$ matrix M over $\mathbb{Z}_m$ and for lists $T \subseteq [t], S \subseteq [s]$ the $T \times S$ submatrix of M is the matrix with rows in T and columns in S. For $i \in [s]$ and $j \in [t]$ we denote the i'th row of M by $M(i :)$ and the j'th column by $M(: j)$.

Definition 4.1 (Span of a set)**.** *For $A \subseteq \mathbb{Z}_m^n$ let* $\operatorname{span}(A)$ *denote the additive subgroup generated by A. We say that a set A spans $u \in \mathbb{Z}_m^n$ if $u \in \operatorname{span}(A)$.*

Definition 4.2 (Rank of a matrix over $\mathbb{Z}_m$)**.** *Let M be a $t \times t$ matrix over $\mathbb{Z}_m$. Then* $\operatorname{rank}(M)$ *is the smallest r such that $M = AB$ where A is an $t \times r$ martrix over $\mathbb{Z}_m$ and B is an $r \times t$ matrix over $\mathbb{Z}_m$.*

Definition 4.3 (Column rank of a matrix over $\mathbb{Z}_m$)**.** *Let M be a $t \times t$ matrix over $\mathbb{Z}_m$. Let* $\operatorname{colspan}(M)$ *denote the subgroup of $\mathbb{Z}_m^t$ generated by the columns of M. The column rank of M over $\mathbb{Z}_m$ is defined as*

$$\operatorname{colrank}(M) = \log_m |\operatorname{colspan}(M)| .$$

The column rank is, in general, a real number in the range $[0, t]$.

Since the rank can behave in unexpected ways over $\mathbb{Z}_m$, we make sure to prove some of the basic facts that we will be using later on.

Fact 4.4. *Let M be a $t \times t$ matrix over $\mathbb{Z}_m$ and let M' be any submatrix of M. Then* $\operatorname{colrank}(M') \leq \operatorname{colrank}(M)$.

Proof. Suppose M' is given by the first t' rows and the first t'' columns of M. We will define an injective map $f : \operatorname{colspan}(M') \to \operatorname{colspan}(M)$. Given any $x \in \operatorname{colspan}(M')$ we can write $x = \sum_{j=1}^{t''} \alpha_j \cdot M'(: j)$ in some fixed way (there might be several choices of α_j). Define $f(x) = \sum_{j=1}^{t''} \alpha_j \cdot M(: j)$. Then, x is clearly the restriction of $f(x)$ to the first t' indices and so the map is injective. $\square$

Fact 4.5. *Let M be a $t \times t$ matrix over $\mathbb{Z}_m$ and let $s|m$. Then* $\operatorname{rank}\left(M^{(s)}\right) \leq \operatorname{rank}(M)$.

Proof. Suppose there exist an $t \times r$ matrix A and an $r \times t$ matrix B over $\mathbb{Z}_m$ such that $M = AB$. Then $M^{(s)} = A^{(s)}B^{(s)}$ and so the rank of $M^{(s)}$ is at most r. $\square$

We will need the following claims relating the rank and the column rank of matrices over $\mathbb{Z}_m$.

Claim 4.6. *Let M be an $t \times t$ matrix over $\mathbb{Z}_m$. Then,*

$$\frac{\operatorname{rank}(M)}{\log m} \leq \operatorname{colrank}(M) \leq \operatorname{rank}(M) .$$

Proof. Let $r = \operatorname{rank}(M)$ and $r' = \operatorname{colrank}(M)$. We first prove that $r' \leq r$. This is equivalent to proving that $|\operatorname{colspan}(M)| \leq m^r$. Let $M = AB$ where A is an $t \times r$ martrix over $\mathbb{Z}_m$ and B is an $r \times t$ matrix over $\mathbb{Z}_m$. Since the columns of M are all in the span of the columns of A we have that the column span of M can contain at most m^r elements.

We now prove that $r' \geq r/(\log m)$ or, equivalently, $|\operatorname{colspan}(M)| \geq 2^r$. Suppose in contradiction that $|\operatorname{colspan}(M)| < 2^r$. Take a minimal spanning set S of $\operatorname{colspan}(M)$ (that is, a set that spans $\operatorname{colspan}(M)$ and such that no proper subset of it does). Suppose $|S| \geq r$ and consider all linear combinations (over $\mathbb{Z}_m$) of elements of S with coefficients in $\{0,1\} \subseteq \mathbb{Z}_m$. Since $|\operatorname{colspan}(M)| < 2^r$ there are two distinct $0-1$ linear combinations that map to the same element. This means that there is a linear combination with coefficients in $\{1,-1\}$ of the elements of S that is equal to zero. Since both 1 and -1 are invertible modulo m we can write one of the elements of S as a linear combination of the other elements. This contradicts the minimality of S and so, we must have $|S| < r$. This implies that $\operatorname{rank}(M) < r$, a contradiction, since we can write M as the product of the matrix with columns in S with the matrix of coefficients giving the columns of M. □

Claim 4.7. *Let M be an $t \times t$ matrix over $\mathbb{Z}_m$, let $r = \operatorname{rank}(M)$. There exists r' columns of M that span the rest of $M's$ columns such that $r' \leq r \log m$.*

Proof. Take a minimal spanning set S of the columns of M (that is, a set that spans all other columns and such that no proper subset of it spans all columns). If $2^{|S|} > m^r$, then $2^{|S|} > |\operatorname{colspan}(M)|$ (by Claim 4.6) and we proceed as in the proof from Claim 4.6 above. If we look at all the $0-1$ combinations of the columns of S, then there are two distinct $0-1$ linear combinations of the columns that map to the same element of $\operatorname{colspan}(M)$. Thus, let $\sum_i \alpha_i S(:i) = \sum_i \beta_i S(:i)$ where $\alpha_i \neq \beta_i$ for at least one i, say i_0. Therefore, we have $\sum_i (\alpha_i - \beta_i) S(:i) = 0$. Note that $(\alpha_{i_0} - \beta_{i_0}) = \pm 1$ and hence is invertible. This lets us write $S(:i_0)$ as a linear combinations of the remaining columns contradicting the minimality of S. Thus, $r' = |S| \leq r \log m$. □

The following claim shows that the column rank behaves similar to rank in terms of subadditivity.

Claim 4.8. *Let A, B be $t \times t$ matrices over $\mathbb{Z}_m$. Then,* $\operatorname{colrank}(A+B) \leq \operatorname{colrank}(A) + \operatorname{colrank}(B)$.

Proof. We show that $|\operatorname{colspan}(A+B)| \leq |\operatorname{colspan}(A)| \, |\operatorname{colspan}(B)|$. Note that $\operatorname{colspan}(A+B) \subseteq \operatorname{colspan}(A) + \operatorname{colspan}(B) \overset{def}{=} \{a+b | a \in \operatorname{colspan}(A), b \in \operatorname{colspan}(B)\}$. Therefore, $|\operatorname{colspan}(A+B)| \leq |\operatorname{colspan}(A) + \operatorname{colspan}(B)| \leq |\operatorname{colspan}(A)| \, |\operatorname{colspan}(B)|$. □

Claim 4.9. *Let M be a $2t \times 2t$ matrix over $\mathbb{Z}_m$, such that*

$$M = \begin{pmatrix} A & 0 \\ \star & B \end{pmatrix}$$

where A, B and $\star$ are $t \times t$ matrices. Then, $\operatorname{colrank}(A) + \operatorname{colrank}(B) \leq \operatorname{colrank}(M)$.

Proof. We show that $|\operatorname{colspan}(A)| \, |\operatorname{colspan}(B)| \leq |\operatorname{colspan}(M)|$. Let $\operatorname{colspan}(A) = R_1$, $\operatorname{colspan}(B) = R_2$, $\operatorname{colspan}(M) = R$. We define $f : R_1 \times R_2 \to R$ and show that f is injective. Given $r_1 \in R_1$ and $r_2 \in R_2$, let $\alpha_1, \cdots \alpha_t$ and $\beta_1, \cdots \beta_t$ denote coefficients for linear combinations of the columns of A and B respectively that give r_1 and r_2. There might be many such linear combinations but we fix one for each r_i. Then, $f(r_1, r_2) = \sum_{i=1}^{t} \alpha_i M(: i) + \sum_{i=t+1}^{2t} \beta_{i-t} M(: i)$. Now, given a column vector $f(r_1, r_2) \in R$, we uniquely identify r_1 and r_2 as follows. We look at the first t rows and call it s_1. Now $s_1 = r_1$ and let $\alpha_1, \cdots \alpha_t$ be the linear combination fixed for r_1 while defining f. Now, consider $f(r_1, r_2) - \sum_{i=1}^{t} \alpha_i M(: i)$ and call the last t rows s_2. Note that $s_2 = r_2$. □

Claim 4.10. *Let M be a $t \times t$ square matrix over $\mathbb{Z}_m$ with zero diagonal entries. If for some $s|m$, $\operatorname{colrank}\left(M^{(s)}\right) \leq 2$, then there exists at least $t' = t/m^2$ indices such that M restricted to those indices as rows and columns is the all zero matrix modulo s.*

Proof. As $\operatorname{colrank}\left(M^{(s)}\right) \leq 2$, it follows that $\left|\operatorname{colspan}\left(M^{(s)}\right)\right| \leq s^2 \leq m^2$. Hence, $M^{(s)}$ has at most m^2 distinct columns. Therefore, there exists a set of indices S of size $t' \geq t/m^2$ with $S = \{r_1, r_2, \cdots r_{t'}\}$ such that all the columns $M^{(s)}(: r_i)$ are identical. Also, as the diagonal elements are zero modulo m, they are zero modulo s. Thus, the $S \times S$ submatrix is the all zero matrix modulo s. □

5 Collision-Free MV families

In the proof of Theorem 2 it will be useful to assume that the elements of the MV family do not 'collide' when reduced modulo an integer s dividing m. In this section we develop the necessary machinery to allow for this assumption. We start by defining a collision free matching vector family.

Definition 5.1 (Collision free MV family)**.** *A* collision free matching vector family (U, V) *in* $\mathbb{Z}_m^n$ *is a matching vector family such that for all $s|m, s \geq 2$, all elements of U are distinct modulo s, and all elements of V are distinct modulo s. Note that if (U, V) is a collision free matching vector family, then so is any $(U', V') \subseteq (U, V)$.*

Lemma 5.2. *Let $m \geq 2$ be an arbitrary integer. Let s be a divisor of m, such that $1 < s < m$. Let (U, V) be a matching vector family in $\mathbb{Z}_m^n$ such that $\langle u, v \rangle = 0 \pmod{s}$ for all $u \in U, v \in V$. Then, $|(U, V)| \leq \mathbf{MV}(m/s, n \log m)$.*

Proof. Let $U = (u_1, u_2, \cdots u_t)$ and $V = (v_1, v_2, \cdots v_t)$. Recall that $P_{U,V}$ is the inner product matrix. We shall write $P_{U,V}$ as P in the rest of the proof for brevity. Let $r = \operatorname{rank}(P) \leq n$. Hence, by Claim 4.7, there exists $r' \leq r \cdot \log m$ columns of P which span all the columns of P. As each entry of P is a multiple of s we can define a matrix P' over $\mathbb{Z}_{m/s}$ by $P' = (1/s) P$. We have

- $P'_{i,i} = 0 \quad \forall i.$
- $P'_{i,j} \neq 0 \quad \forall i \neq j.$

We next show that the r' columns that span the columns of P also span the columns in P'. Without loss of generality, let the first r' columns of P span the remaining columns of P. For any column j, let $P(: j) = \sum_{i=1}^{r'} c_i P(: i) \pmod{m}$. Since all entries of P are divisible by s, we can divide the expression by s and obtain that $P'(: j) = \sum_{i=1}^{r'} c_i P'(: i) \pmod{m/s}$. Hence, we deduce that $r_{P'} = \operatorname{rank}(P') \leq r' \leq r \log m \leq n \log m$. This implies that $P' = AB$ for some $t \times r_{P'}$ matrix A and some $r_{P'} \times t$ matrix B over $\mathbb{Z}_{m/s}$. Thus, the rows of A and the columns of B form a matching vector family in $\mathbb{Z}_{m/s}^{r_{P'}}$. Therefore, $t \leq \mathbf{MV}(m/s, n \log m)$ as claimed. □

Lemma 5.3 (Bucket Lemma)**.** *For any m, let (U, V) be a matching vector family in $\mathbb{Z}_m^n$. Let $1 < s < m$ be any divisor of m. Then, for any $w \in \mathbb{Z}_s^n$, $|B_s(w, U)| \leq \mathbf{MV}(m/s, n \log m)$. By symmetry, $|B_s(w, V)| \leq \mathbf{MV}(m/s, n \log m)$.*

Proof. We prove that $|B_s(w, U)| \leq \mathbf{MV}(m/s, n)$. For $U = (u_1, u_2, \cdots u_t)$, consider any bucket $B_s(w, U) = U'$ (say). Let $U' = \left(u_{j_1}, u_{j_2}, \cdots u_{j_{t'}}\right)$ where $1 \leq j_1 < j_2 < \cdots j_{t'} \leq t$. Let $V' = \left(v_{j_1}, v_{j_2}, \cdots v_{j_{t'}}\right)$. Now, for any $l, m \in [t']$, $\langle u_{j_l}, v_{j_l}\rangle = 0\,(mod\ m)$. Therefore, $\langle u_{j_m}, v_{j_l}\rangle = 0\,(mod\ s)$. By Lemma 5.2 on (U', V'), $t' \leq \mathbf{MV}(m/s, n \log m)$. □

We use the above lemma repeatedly to obtain a collision free matching vector family.

Lemma 5.4. *Let $m \geq 2$ be any positive integer. Suppose there is a matching vector family (U, V) in $\mathbb{Z}_m^n$. Then, there exists a collision free matching vector family $(U', V') \subseteq (U, V)$ such that*

$$\left|(U', V')\right| \geq \frac{|(U, V)|}{\left(\prod_{s|m, 1<s<m} \mathbf{MV}(s, n \log m)\right)^2}.$$

Proof. We will get rid of collisions iteratively by repeatedly applying Lemma 5.3. Let us write the divisors of m in ascending order as $2 \leq s_1 < s_2 < \cdots < s_l \leq m/2$. Perform the following operation for each $s|m$ starting from the smallest divisor greater than 1. For $0 \leq i \leq l$, let U_i, V_i be the matching vector after stage i with $U_0 = U$ and $V_0 = V$. Now suppose that we have U_i, V_i after the i'th stage such that there is no collision modulo s_j in U_i for $1 \leq j \leq i$. The $(i+1)$'th stage is performed as follows. Let us construct U_{i+1}, V_{i+1} from U_i, V_i to ensure no collision among the elements of U_{i+1} modulo s_{i+1} as well. For each $w \in \mathbb{Z}_{s_{i+1}}^n$, by Lemma 5.3, $\left|B_{s_{i+1}}(w, U_i)\right| \leq \mathbf{MV}(m/s_{i+1}, n \log m)$. Pick one element from each bucket in U_i and the corresponding matching vector from V_i to form (U_{i+1}, V_{i+1}). Thus, $|(U_{i+1}, V_{i+1})| \geq |U_i| / \mathbf{MV}(m/s_{i+1}, n \log m)$. We end up with matching vector family U_l, V_l such that $|(U_l, V_l)| \geq \frac{|(U,V)|}{\prod_{s|m,1<s<m} \mathbf{MV}(m/s, n \log m)}$ and U_l is collision free. We repeat the same process this time pruning V_l in order to make it collision free as well. Thus, eventually we end up with a collision free matching vector family $(U_l', V_l') \subseteq (U, V)$ such that

$$\left|(U_l', V_l')\right| \geq \frac{|(U, V)|}{\left(\prod_{s|m, 1<s<m} \mathbf{MV}(m/s, n \log m)\right)^2} = \frac{|(U, V)|}{\left(\prod_{s|m, 1<s<m} \mathbf{MV}(s, n \log m)\right)^2}.$$

□

6 Proof of Theorem 2

Before proceeding with the proof we give yet another definition.

Definition 6.1. *Let $A, B \subseteq \mathbb{Z}_m^n$ be twin-free lists (or sets). Let ω be a primitive root of unity of order m. The duality measure of A, B with respect to ω is defined as*

$$D_\omega(A, B) = \left|\mathbb{E}_{a \sim A, b \sim B}\left[\omega^{\langle a, b\rangle}\right]\right|.$$

Notice that, if $\omega \neq 1$, $D_\omega(A,B) = 1$ implies that there is some $c \in \mathbb{Z}_m$ such that all the entries of the inner product matrix $P_{A,B}$ equal c. We often refer to such submatrices as monochromatic rectangles.

The following is an easy consequence of Lemma 2.13.

Lemma 6.2. *Let (U,V) be an MV family in $\mathbb{Z}_m^n$ of size $t \geq 3m$ and let $\omega = exp(2\pi i/m)$ be a primitive root of unity of order m. Then there exists some $1 \leq j \leq m-1$ such that*

$$D_{\omega^j}(U,V) \geq \frac{2}{3m^{3/2}}.$$

Proof. Let μ be the random variable which chooses $u \in U$ and $v \in V$ randomly and outputs $\langle u, v\rangle$ and let $\mathcal{U}_m$ be the uniform distribution over $\mathbb{Z}_m$. Now, $\Delta(\mu, \mathcal{U}_m) \geq (1/2)(\mathbf{Pr}[\mathcal{U}_m = 0] - \mathbf{Pr}[\mu = 0]) = (1/2)(1/m - 1/t) \geq 1/3m$ as $t \geq 3m$. By Lemma 2.13, for some $1 \leq j \leq m-1$,

$$\left|\mathbb{E}_{x\sim\mu}\left[\left(\omega^j\right)^x\right]\right| \geq \frac{2}{3m^{3/2}}.$$

Thus, we have $\left|\mathbb{E}_{u\sim U, v\sim V}\left[\left(\omega^j\right)^{\langle u,v\rangle}\right]\right| \geq \frac{2}{3m^{3/2}}$ as claimed. □

An important ingredient in the proof of Theorem 2 is the following lemma, referred to in the introduction as the 'sub-matrix lemma' which is a generalization of a result of [BSLZ11].

Lemma 6.3 (Sub-Matrix Lemma). *Let $s, m, n \geq 2$ where s divides m, and let ω be a primitive root of unity of order s. Let $A, B \subset \mathbb{Z}_s^n$ be two twin-free lists satisfying $D_\omega(A,B) \geq \frac{2}{3m^{3/2}}$. Let* $\operatorname{rank}(P_{A,B}) = r \geq 2$. *Then assuming Conjecture 1 (PFR conjecture), there exist lists $A' \subseteq A, B' \subseteq B$ such that $D_\omega(A', B') = 1$, where $|A'| \geq 2^{-c(m)r/\log r}|A|$, $|B'| \geq 2^{-c(m)r/\log r}|B|$ for some constant $c(m)$ which depends only on m.*

Without loss of generality, we can assume $c(m) \geq 1$ above (it will be convenient to assume it in the proof of Theorem 2). In other words, we can replace the $c(m)$ above by $\max\{c(m), 1\}$. We postpone the proof of Lemma 6.3 to Section 7 and proceed now with the proof of Theorem 2.

We restate Theorem 2 here for convenience and with the explicit function $d(m)$.

Theorem 6.4. *Let $n, m \geq 2$ be arbitrary positive integers. Then, assuming Conjecture 1 (PFR conjecture), we have*

$$\mathbf{MV}(m,n) < 2^{d(m)n/\log n},$$

where $d(m) = 1200c(m)m^{6\log m}$ and $c(m)$ is as in Lemma 6.3.

Proof. We prove the theorem by induction on the number of (not necessarily distinct) prime factors of m.

Choice of $d(m)$. Let $d, d_1, d_2, d_3 : \mathbb{Z}^+ \to \mathbb{R}$ be functions and d_4 be a constant. We want the following conditions to be satisfied for all $m, n \geq 2$.

1. $d(m), d_1(m), d_2(m), d_3(m)$ are monotonically increasing in m

2. $(2n)^m \leq 2^{d(m)n/\log n}$

3. $(2m)^m \leq 2^{d(m)n/\log n}$

4. $d(m) \geq d(m/2) \cdot 4m \log m$

5. $-d_2(m) + (1/2)\, d(m) > d(m/2) \log m$

6. $2^{(1/2)d(m)n/\log n} \geq 3m 2^{d_2(m)n/\log n}$

7. $d_2(m)\, n/\log n \geq 2 \log m + d_3(m)\, n/\log n$

8. $d_3(m) \geq d_1(m) \cdot d_4 \cdot m \log m$

9. $d_4 \geq 300$

10. $d_1(m) \geq 2c(m)$

11. $d_2 \geq d_3 + 1$

It can be verified that the following choice for the functions meets the above conditions.

- $d(m) = 1200 \cdot c(m) \cdot m^{6 \log m}$
- $d_1(m) = 2 \cdot c(m)$
- $d_2(m) = 602 \cdot c(m) \cdot m \log m$
- $d_3(m) = 600 \cdot c(m) \cdot m \log m$
- $d_4 = 300$

We shall explicitly mention which conditions of the above functions are being used in different parts of the proof.

Base Case. The base case is where $m = p$ is prime. Lemma 2.9 implies that $\mathbf{MV}(p, n) \leq 1 + \binom{n+p-2}{p-1} < (2 \max\{n, p\})^p$. If we show $(2n)^p \leq 2^{d(p)n/\log n}$ and $(2p)^p \leq 2^{d(p)n/\log n}$ we will be done. Indeed, by the choice of $d(m)$ (Condition 2 and 3) both of the above will hold.

Inductive Case. Let $n \geq 2, m \geq 2$ be arbitrary positive integers. Suppose, by induction, that $\mathbf{MV}(s, n) < 2^{d(s)n/\log n}$ for all $s|m, s < m$. We need to show that, assuming Conjecture 1,

$$\mathbf{MV}(m, n) < 2^{d(m)n/\log n}$$

Suppose not. That is, there exists a matching vector family (U, V) of size $t \geq 2^{d(m)n/\log n}$. First, we shall apply Lemma 5.4 to (U, V) to obtain a large enough collision free matching vector family (U', V').

A large collision free matching vector family. We show that $|(U',V')| \geq 2^{(1/2)d(m)n/\log n}$. Let $|(U',V')| = t'$. Observe that by Lemma 5.4, the inductive hypothesis and the monotonicity of $d(m)$ (Condition 1), $t' \geq 2^{d(m)n/\log n - 2m \cdot d(m/2) \cdot n \log m/\log n}$ where we have used a loose upper bound of m for the number of factors of m. Now,

$$\begin{aligned} & t' \geq 2^{(1/2)d(m)n/\log n} \\ if \quad & d(m)\,n/\log n - 2m \cdot d(m/2) \cdot n \log m/\log n \geq (1/2)\,d(m)\,n/\log n \\ \Leftrightarrow \quad & d(m) \geq d(m/2) \cdot 4m \log m \end{aligned}$$

which is satisfied by the choice of $d(m)$ (Condition 4).

Two key claims. We will need two claims from which the inductive claim follows easily. We shall provide proofs to these claims after the proof of the inductive claim.

Claim 6.5. *Let (U,V) be a collision free matching vector family in $\mathbb{Z}_m^n$ with $|(U,V)| \geq 3m$ and $\operatorname{colrank}\left(P_{U,V}^{(s')}\right) > 2$ for all $s'|m, s' \geq 2$. Then, for some $s|m, s \geq 2$, there exists a collision free matching vector family $(U',V') \subseteq (U,V)$ in $\mathbb{Z}_m^n$ satisfying*

- $|(U',V')| \geq 2^{-d_1(m)r_s/\log r_s}\,|(U,V)|$ *where* $r_s = \operatorname{rank}\left(P_{U,V}^{(s)}\right)$.
- *Either* $\operatorname{colrank}\left(P_{U',V'}^{(s)}\right) \leq (3/4)\operatorname{colrank}\left(P_{U,V}^{(s)}\right)$ *or* $\operatorname{colrank}\left(P_{U',V'}^{(s)}\right) \leq 2$.

Claim 6.6. *Let (U,V) be a collision free matching vector family in $\mathbb{Z}_m^n$ such that $|(U,V)| \geq 3m \cdot 2^{d_2(m)n/\log n}$. Then, there exists a collision free matching vector family $(U',V') \subseteq (U,V)$ in $\mathbb{Z}_m^n$ satisfying*

- $|(U',V')| \geq 2^{-d_2(m)n/\log n}\,|(U,V)|$.
- $P_{U,V}^{(s)}$ *is the all zero matrix for some $s|m, s \geq 2$.*

Let us proceed with the proof of the inductive claim assuming these two claims. We have a collision free matching vector family (U',V') with $|(U',V')| \geq 2^{(1/2)d(m)n/\log n} \geq 3m \cdot 2^{d_2(m)n/\log n}$. (Condition 6 satisfied by the choice of $d(m), d_2(m)$) Applying Claim 6.6, there exists a collision free matching vector family $(U'',V'') \subseteq (U',V') \subseteq (U,V)$ in $\mathbb{Z}_m^n$ satisfying

- $|(U'',V'')| \geq 2^{-d_2(m)n/\log n} 2^{(1/2)d(m)n/\log n}$.
- $P_{U'',V''}^{(s)}$ is the all zero matrix for some $s|m, s \geq 2$.

By the choice of $d(m)$, it can be verified that $-d_2(m) + (1/2)\,d(m) > d(m/2)\log m$ (Condition 5). Thus, $|(U'',V'')| > 2^{d(m/2)n\log m/\log n}$.

We now show that this is enough to get a contradiction. If $s = m$, we have $|(U'',V'')| \leq 1$ as (U'',V'') is a matching vector family in $\mathbb{Z}_m^n$. If $s < m$, by Lemma 5.2 and the inductive hypothesis, we have $|(U'',V'')| \leq 2^{d(m/s)n\log m/\log(n\log m)} \leq 2^{d(m/2)n\log m/\log n}$ by monotonicity of $d(m)$ (Condition 1). Thus, irrespective of s, $|(U'',V'')| \leq 2^{d(m/2)n\log m/\log n}$ which is a contradiction. This completes the proof. □

Proof of Claim 6.5: Let $|(U,V)| = t \geq 3m$. Let ω be a root of unity of order m. By Lemma 6.2, for some $1 \leq j \leq m-1$, $D_{\omega^j}(U,V) \geq \frac{2}{3m^{3/2}}$. Note that $s = m/gcd(m,j)$ is the order of $\omega' = \omega^j$. Observe that $s|m, s \geq 2$ as $1 \leq j \leq m-1$. Recall from the statement of the claim that $r_s = rank\left(P^{(s)}_{U,V}\right)$. Thus, by the collision free property of (U,V),

$$D_{\omega'}\left(U^{(s)}, V^{(s)}\right) = \left|\mathbb{E}_{u \sim U^{(s)}, v \sim V^{(s)}}\left[(\omega')^{\langle u,v\rangle}\right]\right| = \left|\mathbb{E}_{u \sim U, v \sim V}\left[(\omega')^{\langle u,v\rangle}\right]\right| = D_{\omega'}(U,V) \geq \frac{2}{3m^{3/2}}.$$

Applying Lemma 6.3 on $U^{(s)}, V^{(s)}$ with ω' a primitive root of unity of order s, we can get an $(R \times S)$ submatrix of $P_{U,V}$ with $|R| = |S| \geq 2^{-c(m)r_s/\log r_s}t$. (we can make $|R| = |S|$ as throwing away rows and columns from a monochromatic rectangle still keeps it monochromatic) Let $T = R \cap S$. We divide our analysis to two cases: either $|T| > |R|/2$ or $|T| \leq |R|/2$. In both cases, we shall exhibit a matching vector family as required in the statement of the claim.

Case 1: $|T| > |R|/2$. For $U = (u_1, u_2, \cdots u_t)$, $V = (v_1, v_2, \cdots v_t)$, let $U' = (u_j | j \in T)$ and $V' = (v_j | j \in T)$, and $P' = P_{U',V'}$. Now, as $P'^{(s)}$ is monochromatic, and $\langle u_j, v_j\rangle = 0 \ (mod\ s)$ for $j \in T$, we have $\langle u', v'\rangle = 0\,(mod\ s)$ for all $u' \in U', v' \in V'$. Observe that

- $|(U',V')| \geq 2^{-1-c(m)r_s/\log r_s}t \geq 2^{-2c(m)r_s/\log r_s}t \geq 2^{-d_1(m)r_s/\log r_s}t$ (by the choice of $d_1(m)$, Condition 10)
- colrank $\left(P^{(s)}_{U',V'}\right) = 0 \leq 2$

This finishes Case 1.

Case 2: $|T| \leq |R|/2$. Let $R' = R \setminus T$ and $S' = S \setminus T$. Note that $R' \cap S' = \emptyset$ and $|R'| = |S'|$. Consider the $R' \cup S' \times R' \cup S'$ submatrix of $P_{U,V}$. Call it P'. Note that

$$P'^{(s)} = \begin{pmatrix} P'_1 & C \\ \star & P'_2 \end{pmatrix}$$

where P'_1 and P'_2 are the $R' \times R'$ and the $S' \times S'$ submatrices of $P^{(s)}_{U,V}$ respectively and C is monochromatic. We add a matrix of column rank at most 1 to $P'^{(s)}$ to yield $P''^{(s)}$ which is the same as $P'^{(s)}$ except that C is replaced by the all zero block matrix. Thus,

$$P''^{(s)} = \begin{pmatrix} P'_1 & 0 \\ \star & P'_2 \end{pmatrix}$$

Note that by Claim 4.8, colrank $\left(P''^{(s)}\right) \leq$ colrank $\left(P'^{(s)}\right) + 1$. Now, using Claim 4.9, colrank $(P'_1) +$ colrank $(P'_2) \leq$ colrank $\left(P'^{(s)}\right) + 1 \leq$ colrank $\left(P^{(s)}_{U,V}\right) + 1 \leq (3/2)$ colrank $\left(P^{(s)}_{U,V}\right)$ as colrank $\left(P^{(s)}_{U,V}\right) > 2$. Therefore, one of P'_1, P'_2, say P'_1 satisfies colrank $(P'_1) \leq (3/4)$ colrank $\left(P^{(s)}_{U,V}\right)$. Construct the matching vector family (U', V') as follows. Let $U' = (u_j | j \in R')$ and $V' = (v_j | j \in R')$. Again, observe that

- $|(U',V')| \geq 2^{-1-c(m)r_s/\log r_s}t \geq 2^{-2c(m)r_s/\log r_s}t \geq 2^{-d_1(m)r_s/\log r_s}t$ (by the choice of $d_1(m)$, Condition 10).

- colrank $\left(P^{(s)}_{U',V'}\right) \leq (3/4)$colrank $\left(P^{(s)}_{U,V}\right)$.

This completes the proof of Case 2.

□

Proof of Claim 6.6: We will use Claim 6.5 iteratively. For this, we first set up some notations.

The setup. Define a sequence of collision free matching vector families for $i = 0, \ldots, z$.

- $(U, V) = (U_0, V_0), (U_1, V_1) \cdots$
- Let $t_i = |(U_i, V_i)|$.
- Each step i has label $s_i|m$ (this label will be given by Claim 6.5).
- Let $cr_i : \mathbb{Z}^+ \to \mathbb{R}$ be defined by
$$cr_i(s) = \text{colrank}\left(P^{(s)}_{U_i,V_i}\right).$$
- Let $r_i : \mathbb{Z}^+ \to \mathbb{Z}$ be defined by
$$r_i(s) = \text{rank}\left(P^{(s)}_{U_i,V_i}\right).$$

Invariants. We will show how to go from step i to step $i+1$. We stop after stage z when $cr_z(s) \leq 2$ for some $s|m, s \geq 2$. We shall maintain the following invariants for $0 \leq i \leq z-1$.

- $(U_{i+1}, V_{i+1}) \subseteq (U_i, V_i)$ and hence is a collision free matching vector family in $\mathbb{Z}^n_m$.
- $t_{i+1} \geq 2^{-d_1(m) r_i(s_i)/\log r_i(s_i)} t_i$.
- $cr_{i+1}(s_i) \leq (3/4)\, cr_i(s_i)$ or $cr_{i+1}(s_i) \leq 2$.
- $cr_{i+1}(s') \leq cr_i(s')$ for all $s'|m$.

Step $i \to$ Step $i+1$. We state a claim that we will prove below.

Claim 6.7. *$\sum_{i=0}^{z-1} d_1(m)\, r_i(s_i) / \log r_i(s_i) \leq d_3(m)\, n/logn$.*

In order to apply Claim 6.5, we need to satisfy $t_i \geq 3m$. Observe that by Claim 6.7,

$$\begin{aligned} t_i \geq t_z &\geq t_0 \prod_{j=0}^{z-1} 2^{-d_1(m) r_j(s_j)/\log r_j(s_j)} \\ &\geq 2^{-d_3(m)n/\log n} t_0 \\ &\geq 3m \cdot 2^{-d_3(m)n/\log n + d_2(m)n/\log n} \geq 3m, \end{aligned}$$

(by the choice of $d_2(m), d_3(m)$ in Condition 11). Apply Claim 6.5 to (U_i, V_i) to get label s_i for step i and $(U_{i+1}, V_{i+1}) \subseteq (U_i, V_i)$. The first three invariants are maintained by the statement of

Claim 6.5. The last invariant follows from Fact 4.4. Note that by the inequality we just established, $t_z \geq 2^{-d_3(m)n/\log n} t_0$. Also, by the stopping condition, $cr_z(s') \leq 2$ for some $s'|m, s' \geq 2$. Thus, applying Claim 4.10, we get another matching vector family $(U', V') \subseteq (U_z, V_z) \subseteq (U, V)$ such that

- $|(U', V')| \geq t_z/m^2 \geq 2^{-2\log m - d_3(m)n/\log n} |(U, V)| \geq 2^{-d_2(m)n/\log n} |(U, V)|$ (Condition 7 satisfied by the choice of $d_2(m)$ and $d_3(m)$).
- $P^{(s')}_{U',V'}$ is the all zero matrix.

This finishes the Proof of Claim 6.6.

Proof of Claim 6.7: Let t_s be the number of steps with label s. Note that as the column rank modulo s goes down by a factor of at least $3/4$ each time we are in a step labeled s, it is easy to see that $t_s \leq \log_{4/3} cr_0(s) \leq \log_{4/3} n$. We shall rely on the monotonic increasing nature of $x/\log x$ when $x \geq e$. As $cr_i(s) > 2$, by Claim 4.6, $r_i(s) \geq cr_i(s) > 2$ which means $r_i(s) \geq 3 > e$ as the rank is always an integer. We thus have

$$\begin{aligned}
& \sum_{i=0}^{z-1} d_1(m) \frac{r_i(s_i)}{\log r_i(s_i)} \\
\leq \; & d_1(m) \log m \sum_{i=0}^{z-1} \frac{cr_i(s_i)}{\log cr_i(s_i)} \quad \text{(by Claim 4.6) and monotonicity of } x/\log x \text{ as discussed above} \\
\leq \; & d_1(m) \log m \sum_{s|m, s\geq 2} \; \sum_{j=1}^{\lfloor \log_{4/3} n(s) \rfloor} \left(\frac{cr_0(s)}{(4/3)^{j-1} \log\left(cr_0(s) / (4/3)^{j-1} \right)} \right) \\
\leq \; & d_1(m) \log m \sum_{s|m, s\geq 2} d_4 cr_0(s) / \log cr_0(s) \quad \text{(by Claim A.1 and Condition 9 satisfied by } d_4\text{)} \\
\leq \; & d_1(m) \log m \sum_{s|m, s\geq 2} d_4 n/\log n \quad \text{(as } cr_0(s) \leq r_0(s) \leq r_0(m) \leq n\text{, by Claim 4.6 and Fact 4.5)} \\
\leq \; & d_4 d_1(m)\, m (\log m)\, n/\log n \\
\leq \; & d_3(m)\, n/\log n \quad \text{(by the choice of } d_3(m)\text{, Condition 8)}
\end{aligned}$$

This completes the proof. □

7 Monochromatic rectangles from low rank matrices

In this section we prove Lemma 6.3 (the Sub-Matrix Lemma). We begin with some preliminary definitions. The following is a standard result in algebra and can be find in any introductory text.

Theorem 7.1 (Fundamental Theorem of finitely generated abelian groups)**.** *Every finitely generated abelian group G is isomorphic to a direct product of cyclic groups of prime power order and an infinite cyclic group. More precisely,*

$$G \cong \mathbb{Z}^n \times \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_2} \cdots \times \mathbb{Z}_{q_r}$$

where q_i's are prime powers with $q_1 \leq q_2 \cdots \leq q_r$. The decomposition is unique after applying this ordering on q_i's. If the group G is finite, then $n = 0$.

We will use the following two definitions regarding sumsets.

Definition 7.2 (Difference Set)**.** *For $A \subseteq \mathbb{Z}_m^n$ define its difference set as $A - A$=$\{a - a' | a, a' \in A\}$.*

Definition 7.3 ($reps_S(x)$)**.** *For any $S \subseteq \mathbb{Z}_m^n$ and $x \in \mathbb{Z}_m^n$, $reps_S(x)$ is the number of different representations of x as an expression of the form $s - s'$ where $s, s' \in S$.*

Next, we define the ϵ-spectrum of B with respect to a primitive root of unity of order m.

Definition 7.4 (Spectrum)**.** *For $B \subseteq \mathbb{Z}_m^n$, and $\epsilon \in [0, 1]$, the ϵ-spectrum of B with respect to ω, a primitive root of unity of order m, is the set*

$$\mathrm{Spec}_\epsilon(B) = \left\{ x \in \mathbb{Z}_m^n : \left| \mathbb{E}_{b \sim B} \left[\omega^{\langle x, b \rangle} \right] \right| \geq \epsilon \right\}.$$

When ω is implicit in the context, we will drop the phrase "with respect to ω".

We start by proving the following lemma which is a generalization of a lemma from [BSLZ11].

Lemma 7.5. *Let $A, B \subseteq \mathbb{Z}_m^n$ be sets. Let ω be a primitive root of unity of order m. If $A \subseteq \mathrm{Spec}_\epsilon(B)$, then there exist sets $A' \subseteq A, B' \subseteq B$, such that $|A'| \geq |A|/m$ and $|B'| \geq \epsilon^2 \frac{|A|}{|span(A)|}|B|$ such that $D_\omega(A', B') = 1$.*

Proof. We start by setting up some notations. Let $W = \mathrm{span}(A)$ be the subgroup of $\mathbb{Z}_m^n$ spanned by A. By Theorem 7.1, there exists an isomorphism $\tau : \prod_{i=1}^r \mathbb{Z}_{q_i} \to W$. Let $\mathcal{C} = \prod_{i=1}^r \mathbb{Z}_{q_i}$ and note that we can think of elements of $\mathcal{C}$ as vectors with integer coordinates where the i'th coordinate is in $\mathbb{Z}_{q_i}$. Let $e_1, e_2, \cdots e_r \in \mathcal{C}$ where e_i is the vector that has 1 in the i'th coordinate and 0 everywhere else. Given $x \in \mathcal{C}$, $\exists \alpha_1, \cdots \alpha_r$, with $\alpha_i \in \mathbb{Z}_{q_i}$ such that

$$x = \sum_{i=1}^r \alpha_i e_i.$$

Then $\tau(x) = \sum_{i=1}^r \alpha_i \tau(e_i)$. Let $v_i = \tau(e_i)$ for $1 \leq i \leq r$. We can think of the v_i's as a basis of W. Therefore, for $\alpha = (\alpha_1, \alpha_2, \cdots \alpha_r) \in \mathcal{C}$ we have $\tau(\alpha) = \sum_{i=1}^r \alpha_i v_i$. Let

$$\Theta = \{(\beta_1, \cdots \beta_r) \in \mathbb{Z}_m^r | \exists u \in \mathbb{Z}_m^n \text{ such that } \forall i, \beta_i = \langle v_i, u \rangle\}.$$

Claim 7.6. *For $1 \leq i \leq r$, $q_i v_i = 0^n \pmod{m}$.*

Proof. Let $x = 0^r \in \mathcal{C}$. Now $\tau(x) = 0^n \pmod{m}$. Note that x can also be written as $x = q_i e_i$. Applying τ on both sides, we get $\tau(x) = q_i v_i$. Thus, $q_i v_i = 0^n \pmod{m}$. □

Claim 7.7. *For $\beta \in \Theta$, $1 \leq i \leq r$, $q_i \beta_i = 0 \pmod{m}$.*

Proof. As $\beta \in \Theta$, there is a $u \in \mathbb{Z}_m^n$ such that $\forall i$, $\beta_i = \langle v_i, u \rangle$. Then, $q_i \beta_i = q_i \langle v_i, u \rangle = 0 \pmod{m}$ by Claim 7.6. □

For $\alpha \in \mathcal{C}, \beta \in \Theta$ we define their inner product $\langle \alpha, \beta \rangle \in \mathbb{Z}_m$ by considering $\alpha_i \in \{0, \ldots, q_i - 1\}, \beta_i \in \{0, \ldots, m-1\}$, taking the inner product over the integers and then reducing the result modulo m. This is indeed an inner product by Claim 7.7.

Claim 7.8. *Given $\beta \in \Theta \setminus \{0\}$,*

$$\sum_{a \in W} \omega^{\langle \tau^{-1}(a), \beta \rangle} = \sum_{\alpha \in \mathcal{C}} \omega^{\langle \alpha, \beta \rangle} = 0.$$

Proof. Let $\beta_i \neq 0$. Then $\sum_{\alpha \in C} \omega^{\langle \alpha, \beta \rangle} = 0$ whenever $\sum_{j=0}^{q_i - 1} \omega^{j \beta_i} = 0$. Now, $\sum_{j=0}^{q_i-1} \omega^{j\beta_i} = \frac{\omega^{q_i \beta_i} - 1}{\omega^{\beta_i} - 1}$. This is well defined because ω is of order m and $\beta_i \neq 0$. The claim now follows from Claim 7.7 which makes the expression zero. □

With the above setup in place, we can now proceed with the proof of Lemma 7.5. For $\beta \in \Theta$, define

$$S_\beta = \{x \in \mathbb{Z}_m^n | \langle v_i, x \rangle = \beta_i, 1 \leq i \leq r\}.$$

Denoting $\mu(\beta) = \mathbf{Pr}_{b \in B}[b \in S_\beta]$, we observe that $\cup_{\beta \in \Theta} (B \cap S_\beta) = B$. Hence, $\sum_{\beta \in \Theta} \mu(\beta) = 1$. For $a \in W$, define $h(a) = \mathbb{E}_{b \in B}\left[\omega^{\langle a, b \rangle}\right]$. If $a = \sum_{i=1}^r \alpha_i v_i$ then

$$\begin{aligned} h(a) &= \mathbb{E}_{b \in B}\left[\omega^{\langle a, b \rangle}\right] \\ &= \mathbb{E}_{b \in B}\left[\omega^{\langle \sum_{i=1}^r \alpha_i v_i, b \rangle}\right] \\ &= \sum_{\beta \in \Theta} \mu(\beta) \omega^{\langle \alpha, \beta \rangle} \\ &= \sum_{\beta \in \Theta} \mu(\beta) \omega^{\langle \tau^{-1}(a), \beta \rangle}. \end{aligned}$$

We will prove upper and lower bounds for the sum $\sum_{a \in A} |h(a)|^2$. On the one hand,

$$\begin{aligned} \sum_{a \in A} |h(a)|^2 &\geq \frac{1}{|A|} \left(\sum_{a \in A} |h(a)| \right)^2 \quad \text{(Cauchy Scwartz inequality)} \\ &\geq \frac{1}{|A|} \left(\sum_{a \in A} \epsilon \right)^2 \quad (A \subseteq \text{Spec}_\epsilon(B) \text{ implies } |h(a)| \geq \epsilon\ ) \\ &\geq |A| \epsilon^2. \end{aligned}$$

On the other hand,

$$\begin{aligned} \sum_{a \in A} |h(a)|^2 &\leq \sum_{a \in W} |h(a)|^2 \\ &= \sum_{a \in W} \sum_{\beta \in \Theta, \beta' \in \Theta'} \mu(\beta) \mu(\beta') \omega^{\langle \tau^{-1}(a), \beta - \beta' \rangle} \\ &= \sum_{\beta, \beta' \in \Theta,} \mu(\beta) \mu(\beta') \sum_{a \in W} \omega^{\langle \tau^{-1}(a), \beta - \beta' \rangle} \\ &= \sum_{\beta \in \Theta} \mu(\beta)^2 |W| \qquad \text{(Claim 7.8)} \\ &\leq |W| \max_{\beta \in \Theta} \{\mu(\beta)\}. \end{aligned}$$

Now, combining the upper and lower bounds, $\max_{\beta\in\Theta}\{\mu(\beta)\} \geq \epsilon^2\frac{|A|}{|W|}$. Thus, there exists a $\beta \in \Theta$ such that $\mu(\beta) \geq \epsilon^2\frac{|A|}{|W|}$. This means that the subset $B' = B \cap S_\beta$ is of size at least $\epsilon^2\frac{|A|}{|W|}|B|$. Now, any $a \in A$ can be written as $a = \sum_{i=1}^{r}\alpha_i v_i$, and for $b \in B'$, the inner product $\langle a, b\rangle = \langle \alpha, \beta\rangle$ is independent of b. Now, for $i \in [m]$, let $A_i \subseteq A$ be such that for $a \in A_i$, for all $b \in B'$, $\langle a, b\rangle = \langle \alpha, \beta\rangle = i$. Now there exists some A_i, call it A', of size at least $|A|/m$ such that $\langle a', b'\rangle = i$ for all $a' \in A', b' \in B'$, that is, $D_\omega(A', B') = 1$ and this proves the lemma. □

We continue along the lines of [BSLZ11] and prove the following lemma.

Lemma 7.9. *Suppose the twin free lists $U, V \subseteq \mathbb{Z}_m^n$ satisfy $D_\omega(U, V) \geq \epsilon$ where ω is a primitive root of unity of order m. Also, let* $\mathrm{rank}(P_{U,V}) = r$. *Then assuming Conjecture 1, for every $K > 1$, letting $\ell = r/\log_m K$, there exist lists $U' \subseteq U, V' \subseteq V$ such that $D_\omega(U', V') = 1$, and $|U'| \geq poly_m\left(\frac{(\epsilon/2)^{2^\ell}}{rK}\right)(2mr)^{-\ell}|U|$, $|V'| \geq poly_m\left(\frac{(\epsilon/2)^{2^\ell}}{rK}\right)m^{-\ell}|V|$.*

Proof. Let $U = (u_1, \cdots u_t)$ and $V = (v_1, \cdots v_t)$. Since $P_{U,V}$ has rank r there exists a $t \times r$ matrix U_M and $r \times t$ matrix V_M so that $U_M V_M = P_{U,V}$. Thus if we let A denote the rows of U_M and B denote the columns of V_M, then $A, B \subseteq \mathbb{Z}_m^r$. The proof does not care about the order of elements and hence we now consider A, B which are sets. Note that $|A| = |B| = t$ and if $A = (a_1, \cdots a_t)$ and $B = (b_1, \cdots b_t)$ then $\langle a_i, b_j\rangle = \langle u_i, v_j\rangle$ for $1 \leq i, j \leq t$. Thus, $D_\omega(U, V) \geq \epsilon$ implies $D_\omega(A, B) \geq \epsilon$. Following [BSLZ11] consider a sequence of constants $\epsilon_1 = \epsilon/2$, $\epsilon_2 = \epsilon_1^2/2$, $\epsilon_3 = \epsilon_2^2/2$, $\cdots$ and a sequence of sets $A_1 = A \cap \mathrm{Spec}_{\epsilon_1}(B)$ and $A_i \subseteq (A_{i-1} - A_{i-1}) \cap \mathrm{Spec}_{\epsilon_i}(B)$. The way the subsets are chosen for A_i's will be made precise shortly. Now by the pigeonhole principle, there exists a minimal index $\ell \leq r/\log_m K$ such that $|A_{\ell+1}| \leq K|A_\ell|$. To give a precise definition of the A_i's , we have the following. Let $A_1 = A \cap \mathrm{Spec}_{\epsilon/2}(B)$. For $i \geq 2$, assuming ϵ_{i-1} and A_{i-1}, let j_i be the the integer index which maximizes the size of

$$\{(a, a') \in A_{i-1} \times A_{i-1} | a - a' \in \mathrm{Spec}_{\epsilon_i}(B) \textit{ and } m^{j_i} \leq rep_{A_{i-1}}(a - a') \leq m^{j_i+1}\},$$

and let

$$A_i = \{a - a' | a, a' \in A_{i-1}, a - a' \in \mathrm{Spec}_{\epsilon_i}(B) \textit{ and } m^{j_i} \leq rep_{A_{i-1}}(a - a') \leq m^{j_i+1}\}.$$

Claim 7.10. *For $i = 1$ we have $|A_1| \geq (\epsilon/2)|A|$. For $i > 1$ we have $\mathbf{Pr}_{a,a'\in A_{i-1}}[a - a' \in A_i] \geq \epsilon_i/r$ and additionally $|A_i| \geq \frac{\epsilon_i}{m^{j_i+1}r}|A_{i-1}|^2$.*

Proof. The case of $i = 1$ follows from Markov inequality. For larger i, we show that

$$\mathbf{Pr}_{a,a'\in A_{i-1}}[a - a' \in \mathrm{Spec}_{\epsilon_i}(B)] \geq \epsilon_i.$$

This follows from the fact that

$$\epsilon_{i-1}^2 \leq \left|\mathbb{E}_{b\in B, a\in A_{i-1}}\left[\omega^{\langle a,b\rangle}\right]\right|^2 \leq \mathbb{E}_{b\in B}\left|\mathbb{E}_{a\in A_{i-1}}\left[\omega^{\langle a,b\rangle}\right]\right|^2 = \mathbb{E}_{a,a'\in A_{i-1}}\mathbb{E}_{b\in B}\left[\omega^{\langle a-a',b\rangle}\right].$$

Now applying Markov inequality we get that $\mathbf{Pr}_{a,a'\in A_{i-1}}[a - a' \in \mathrm{Spec}_{\epsilon_i}(B)] \geq \epsilon_i = \epsilon_{i-1}^2/2$. Now selecting j_i as in the construction gives that $\mathbf{Pr}_{a,a'\in A_{i-1}}[a - a' \in A_i] \geq \epsilon_i/r$.

To prove the second part of the lemma, observe that by the above, we have shown that

$$\left|\{(a,a') \in A_{i-1} \times A_{i-1} | a - a' \in A_i\}\right| \geq \frac{\epsilon_i}{r}|A_{i-1}|^2.$$

Also, by construction of A_i, since every $x \in A_i$ can be represented as $x = a - a'$ with $a, a' \in A_{i-1}$ in at most m^{j_i+1} ways, we have that $|A_i| \geq \frac{\epsilon_i}{m^{j_i+1}r}\,|A_{i-1}|^2$. This completes the proof. □

Below we will use the following additive-combinatorics lemma.

Theorem 7.11 ([BS94, Gow98]). *There exists an absolute constant $c > 0$ such that the following holds. Let A be any arbitrary subset of an abelian group G. Let $S \subseteq G$ be such that $|S| \leq C|A|$. If $\mathbf{Pr}_{a,a' \in A}[a - a' \in S] \geq 1/C$, then there exists a subset $A' \subseteq A$ such that $|A'| \geq \frac{|A|}{C^c}$ and $|A' - A'| \leq C^c|A|$.*

Now we come to the main claim.

Claim 7.12. *For $i = \ell, \ell - 1, \cdots 1$ there exist subsets $A'_i \subseteq A_i$, $B'_i \subseteq B$ such that $D_\omega\left(A'_i, B'_i\right) = 1$ and*

$$|A'_i| \geq \alpha_i|A_i|$$

and

$$|B'_i| \geq \beta_i|B|$$

where $\alpha_i = poly_m\left(\frac{\epsilon_{\ell+1}}{rK}\right)(2mr)^{-(\ell-i)}\left(\prod_{j=i}^{\ell}\epsilon_{j+1}\right)$, $\beta_i = poly_m\left(\frac{\epsilon_{\ell+1}}{rK}\right)m^{-(\ell-i)}$

Base Case. The base case of $i = \ell$ is proved by an application of the Balog-Szemeredi-Gowers theorem followed by Conjecture 1 followed by Lemma 7.5. To see this, we know that $|A_{\ell+1}| \leq K|A_\ell|$ and $\mathbf{Pr}_{a,a' \in A_\ell}[a - a' \in A_{\ell+1}] \geq \epsilon_{\ell+1}/r$. Hence by Theorem 7.11 (with $C = \frac{rK}{\epsilon_{\ell+1}}$), there exists a set $A''_\ell \subseteq A_\ell$ such that $|A''_\ell| \geq poly\left(\frac{\epsilon_{\ell+1}}{rK}\right)|A_\ell|$ and $|A''_\ell - A''_\ell| \leq poly\left(\frac{rK}{\epsilon_{\ell+1}}\right)|A''_\ell|$. Now by Conjecture 1 applied to A''_ℓ, there exists a set $A'''_\ell \subseteq A''_\ell$ such that $|A'''_\ell| \geq poly_m\left(\frac{\epsilon_{\ell+1}}{rK}\right)|A''_\ell|$ and $|\mathrm{span}\,(A'''_\ell)| \leq m|A''_\ell| = poly_m\left(\frac{rK}{\epsilon_{\ell+1}}\right)|A'''_\ell|$. (Note the extra factor of m in front of $|A''_\ell|$ as we get a coset of size $|A''_\ell|$ and its span incurs an additional factor of m) Also, as $A'''_\ell \in \mathrm{Spec}_{\epsilon_\ell}(B)$, applying Lemma 7.5 to A'''_ℓ and B, we get $A'_\ell \subseteq A'''_\ell$ and $B'_\ell \subseteq B$ such that $D_\omega\left(A'_\ell, B'_\ell\right) = 1$, $|A'_\ell| \geq poly_m\left(\frac{\epsilon_{\ell+1}}{rK}\right)|A_\ell|$ and $|B'_\ell| \geq poly_m\left(\frac{\epsilon_{\ell+1}}{rK}\right)|B|$. This completes the base case. Let us come to the inductive case. □

Inductive Case. Suppose the statement is true for i and let us argue for $i-1$. Let $G = (A_{i-1}, E)$ be the graph whose vertices are the elements in A_{i-1} and (a, a') is an edge if $a - a' \in A'_i$. Now,

$$\begin{aligned}
|E| &\geq m^{j_i}|A'_i| \\
&\geq m^{j_i}\alpha_i|A_i| \quad \text{(inductive hypothesis)} \\
&\geq m^{j_i}\alpha_i\frac{\epsilon_i}{m^{j_i+1}r}|A_{i-1}|^2 \quad \text{(Claim 7.10)} \\
&= 2\alpha_{i-1}|A_{i-1}|^2
\end{aligned}$$

Now the graph has at least $2\alpha_{i-1}|A_{i-1}|^2$ edges and $|A_{i-1}|$ vertices and therefore has a connected component of size at least $2\alpha_{i-1}|A_{i-1}|$ vertices. Let us call these vertices A''_{i-1}. Let $\tilde{a}$ be any element of A''_{i-1}. Partition B'_i into $B'_{i,j}$ for $0 \leq j \leq m-1$ such that all elements of $B'_{i,j}$ have inner product j with $\tilde{a}$. Let $B'_{i-1} = B_{i,j_1}$ be the largest of them. Note that $|B'_{i-1}| \geq |B'_i|/m$. By assumption

$D_\omega\left(A'_i, B'_i\right) = 1$. Hence, $D_\omega\left(A'_i, B'_{i-1}\right) = 1$. Therefore, for some j_2, $\langle a, b\rangle = j_2$ for all $a \in A'_i$ and $b \in B'_{i-1}$. Now, in the connected component obtained above, whenever $a, a' \in A''_{i-1}$ are neighbours, $\langle a - a', b\rangle = j_2$ for $b \in B'_{i-1}$. Thus, starting with $\tilde{a}$ as the anchor and propagating throughout the connected component, we can classify the vertices in A''_{i-1} based on the inner product it has with all elements in B'_{i-1}, which is either j_1 or $j_2 - j_1$. Pick the larger set and call it A'_{i-1}. Hence, $D_\omega\left(A'_{i-1}, B'_{i-1}\right) = 1$. Thus, $|A'_{i-1}| \geq |A''_{i-1}|/2 \geq \alpha_{i-1}|A_{i-1}|$ and $B'_{i-1} \geq |B'_i|/m \geq \frac{\beta_i}{m}|B| = \beta_{i-1}|B|$. This completes the inductive case. □

Put $i = 1$ in the above claim. Also observe that as $\epsilon_{j+1} = \epsilon^{2^j}/2^{2^j-1} \geq (\epsilon/2)^{2^j}$. Thus, $\epsilon_{\ell+1} \geq (\epsilon/2)^{2^\ell}$ and $\prod_{j=1}^{\ell} \epsilon_{j+1} \geq (\epsilon/2)^{2^{\ell+1}}$ there exist $A' \subseteq A_1 \subseteq A$, $B' \subseteq B$, such that $|A'| \geq \mathit{poly}\left(\frac{(\epsilon/2)^{2^\ell}}{rK}\right)(2mr)^{-\ell}|A|$ and $|B'| \geq \mathit{poly}\left(\frac{(\epsilon/2)^{2^\ell}}{rK}\right)m^{-\ell}|B|$. Observing that the lower bounds grow weaker with increasing ℓ,and that $\ell \leq \ell' = r/\log_m K$ we get $|A'| \geq \mathit{poly}\left(\frac{(\epsilon/2)^{2^{\ell'}}}{rK}\right)(2mr)^{-\ell'}|A|$ and $|B'| \geq \mathit{poly}\left(\frac{(\epsilon/2)^{2^{\ell'}}}{rK}\right)m^{-\ell'}|B|$ where $\ell' = r/\log_m K$. Therefore, if we take the list $U' \subseteq U$ (corresponding to $A' \subseteq A$) and $V' \subseteq V$ (corresponding to $B' \subseteq B$) then as $\langle a_i, b_j\rangle = \langle u_i, v_j\rangle$ the statement of the lemma follows. This completes the proof of Lemma 7.9 □

We can now prove the Sub-Matrix Lemma, Lemma 6.3.

Proof of Lemma 6.3: Set $K = s^{4r/\log r}, \ell = \frac{\log r}{4}, \epsilon = 1/2m^{3/2}$ while applying Lemma 7.9 over $\mathbb{Z}_s$. We get $|A'| \geq \delta_s|A|$, $|B'| \geq \delta_s|B|$ where

$$\begin{aligned} \delta_s &= \mathit{poly}_s\left(\frac{1}{m^{r^{1/4}}}\right)2^{-c_1(s)r/\log r} \quad \text{(for some constant } c_1(s) \text{ depending only on } s\text{)} \\ &\geq \mathit{poly}_m\left(\frac{1}{m^{r^{1/4}}}\right)2^{-c_1(s)r/\log r} \end{aligned}$$

Now let $c_2(m) = \max_{s|m, s\geq 2}\{c_1(s)\}$. Thus, $\delta_s \geq \mathit{poly}_m\left(\frac{1}{m^{r^{1/4}}}\right)2^{-c_2(m)r/\log r} \geq 2^{-c(m)r/\log r}$ for some constant c that depends only on m. □

References

[BET10] Avraham Ben-Aroya, Klim Efremenko, and Amnon Ta-Shma. Local list decoding with a constant number of queries. In *51st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 715–722, 2010.

[BF98] Laszlo Babai and Peter Frankl. *Linear algebra methods in combinatorics.* 1998.

[BIKR02] Amos Beimel, Yuval Ishai, Eyal Kushilevitz, and Jean-Francios Raymond. Breaking the $O\left(n^{1/(2k-1)}\right)$ barrier for information-theoretic private information retrieval. In *43rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 261–270, 2002.

[BS94] Antal Balog and Endre Szemerédi. A statistical theorem of set addition. *Combinatorica*, 14(3):263–268, 1994.

[BSLZ11] Eli Ben-Sasson, Shachar Lovett, and Noga Zewi. An additive combinatorics approach to the log-rank conjecture in communication complexity. *CoRR*, abs/1111.5884, 2011.

[BSZ11] Eli Ben-Sasson and Noga Zewi. From affine to two-source extractors via approximate duality. In *STOC*, pages 177–186, 2011.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.*, 17(2):230–261, 1988.

[DGY11] Zeev Dvir, Parikshit Gopalan, and Sergey Yekhanin. Matching vector codes. *SIAM J. Comput.*, 40(4):1154–1178, 2011.

[Efr09] Klim Efremenko. 3-query locally decodable codes of subexponential length. In *41st ACM Symposium on Theory of Computing (STOC)*, pages 39–44, 2009.

[GKST02] Oded Goldreich, Howard Karloff, Leonard Schulman, and Luca Trevisan. Lower bounds for locally decodable codes and private information retrieval. In *17th IEEE Computational Complexity Conference (CCC)*, pages 175–183, 2002.

[Gow98] William Timothy Gowers. A new proof of szemer'edis theorem for arithmetic progressions of length four. *Geom. Funct. Anal.*, 17(2):230–261, 1998.

[Gro00] Vince Grolmusz. Superpolynomial size set-systems with restricted intersections mod 6 and explicit Ramsey graphs. *Combinatorica*, 20:71–86, 2000.

[Gro02] Vince Grolmusz. Constructing set-systems with prescribed intersection sizes. *Journal of Algorithms*, 44:321–337, 2002.

[IS10] Toshiya Itoh and Yasuhiro Suzuki. New constructions for query-efficient locally decodable codes of subexponential length. *IEICE Transactions on Information and Systems*, pages 263–270, 2010.

[KdW04] Iordanis Kerenidis and Ronald de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. *Journal of Computer and System Sciences*, 69:395–420, 2004.

[KT00] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *32nd ACM Symposium on Theory of Computing (STOC)*, pages 80–86, 2000.

[KY09] Kiran S. Kedlaya and Sergey Yekhanin. Locally decodable codes from nice subsets of finite fields and prime factors of Mersenne numbers. *SIAM Journal on Computing*, 38:1952–1969, 2009.

[Lov12] Shachar Lovett. An exposition of sanders' quasi-polynomial freiamn-ruzsa theorem. To appear., 2012.

[MFL$^+$10] Yeow Meng Chee, Tao Feng, San Ling, Huaxiong Wang, and Liangfeng Zhang. Query-efficient locally decodable codes of subexponential length. In *Electronic Colloquium on Computational Complexity (ECCC)*, TR10-173, 2010.

[Rag07] Prasad Raghavendra. A note on Yekhanin's locally decodable codes. In *Electronic Colloquium on Computational Complexity (ECCC)*, TR07-016, 2007.

[San10] T. Sanders. On the Bogolyubov-Ruzsa lemma. *ArXiv e-prints*, October 2010.

[Sga99] Jiri Sgall. Bounds on pairs of families with restricted intersections. *Combinatorica*, 19:555–566, 1999.

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

[Woo07] David Woodruff. New lower bounds for general locally decodable codes. In *Electronic Colloquium on Computational Complexity (ECCC)*, TR07-006, 2007.

[Woo10] David P. Woodruff. A quadratic lower bound for three-query linear locally decodable codes over any field. In *Proceedings of the 13th international conference on Approximation, and 14 the International conference on Randomization, and combinatorial optimization: algorithms and techniques*, APPROX/RANDOM'10, pages 766–779, Berlin, Heidelberg, 2010. Springer-Verlag.

[Yek08] Sergey Yekhanin. Towards 3-query locally decodable codes of subexponential length. *Journal of the ACM*, 55:1–16, 2008.

[Yek11] Sergey Yekhanin. Locally decodable codes. *Foundations and trends in theoretical computer science*, 2011. To appear. Preliminary version available for download at http://research.microsoft.com/en-us/um/people/yekhanin/Papers/LDC_now.pdf.

[YGK12] Chen Yuan, Qian Guo, and Haibin Kan. A novel elementary construction of matching vectors. *Information Processing Letters*, 112(12):494 – 496, 2012.

A A Calculation

Claim A.1. *Let $b > 1, n \geq 2$ be arbitrary integers. Then*

$$\sum_{i=1}^{\lfloor \log_b n \rfloor} \frac{1}{b^{i-1} \log{(n/b^{i-1})}} \leq f(b)/\log n$$

where $f(b) = \frac{10b}{b-1} + \frac{10}{\log b} + \frac{16e}{\log^2 b}$. When $b = 4/3$, $f(b) < 300$.

Proof. We divide the summation into two parts. The first part consists of the first $\lfloor \log_b \log n \rfloor$ terms and the second part consists of the remaining terms.

In the first part, $\frac{1}{b^{i-1} \log n/b^{i-1}} \leq \frac{1}{b^{i-1} 0.1 \log n}$ whenever $n \geq 2$ and hence the first part summation is bounded from above by $\frac{10b}{(b-1)\log n}$.

In the second part of the summation, we use the monotonicity of $x \log{(n/x)}$. The function increases with x as long as $x \leq n/e$. Therefore, for terms with $b^{i-1} \leq n/e$, the maximum value of each summand is given by substituting $i = \log_b \log n$ which gives an upper bound of $\frac{1}{0.1 \log^2 n}$.

The remaining terms corresponding to $n/b \geq b^{i-1} > n/e$ (note that these extra terms arise only if $b < e$) can be analysed as follows. Observe that each summand in that range can be upperbounded by $\frac{e}{n \log b}$. Therefore, we have at most $\log_b n$ terms each at most $\frac{10}{\log^2 n} + \frac{e}{n \log b}$. Thus, the second part of the summation is bounded from above by $\log_b n \left(\frac{10}{\log^2 n} + \frac{e}{n \log b} \right)$.

$$\begin{aligned}
\log_b n \left(\frac{10}{\log^2 n} + \frac{e}{n \log b} \right) &= \frac{10}{\log b} \frac{1}{\log n} + \frac{e}{\log^2 b} \frac{\log n}{n} \\
&\leq \frac{10}{\log b} \frac{1}{\log n} + \frac{e}{\log^2 b} \frac{16}{\log n} \quad (\text{as } 16n \geq \log^2 n) \\
&= \left(\frac{10}{\log b} + \frac{16e}{\log^2 b} \right) \frac{1}{\log n}
\end{aligned}$$

This completes the proof. □

B Proofs of Two Probability Lemmas

B.1 Proof of Lemma 2.13

Let $f : \mathbb{Z}_m \to \mathbb{C}$ be any function. Recall that, for $0 \leq j \leq m-1$, the Fourier coefficients of f are given by

$$\hat{f}(j) = \frac{1}{m} \sum_{x \in \mathbb{Z}_m} f(x) \exp(-2\pi i j x/m).$$

It is well known that the set of functions $\{\exp(2\pi i j x/m)\}_{0 \leq j \leq m-1}$ is an orthonormal basis for all functions of the above form, and that f can be expressed as

$$f(x) = \sum_{j=0}^{m-1} \hat{f}(j) \exp(2\pi i j x/m).$$

Let us consider $f : \mathbb{Z}_m \to [0,1]$. Thus, Parseval's identity states that

$$\sum_{j=0}^{m-1} \left| \hat{f}(j) \right|^2 = \frac{1}{m} \sum_{x \in \mathbb{Z}_m} f(x)^2 \leq 1.$$

Observe that as $\mathcal{U}_m(x) = 1/m$ is the constant function, $\hat{\mathcal{U}_m}(j) = 0$ for $j \neq 0$. Also, for any distribution μ, $\hat{\mu}(0) = 1/m$. Now

$$\begin{aligned}
2\epsilon &\leq \sum_{x \in \mathbb{Z}_m} |\mu(x) - \mathcal{U}_m(x)| \\
&\leq \sqrt{m}\sqrt{\sum_{x \in \mathbb{Z}_m} |\mu(x) - \mathcal{U}_m(x)|^2} \qquad \text{(Cauchy Schwartz Inequality)} \\
&= m\sqrt{\sum_{i=0}^{m-1} \left|\left(\hat{\mu}(i) - \hat{\mathcal{U}_m}(i)\right)\right|^2} \\
&= m\sqrt{\sum_{i=1}^{m-1} |\hat{\mu}(i)|^2} \qquad (\hat{\mathcal{U}_m}(j) = 0 \text{ for } j \neq 0\text{, and } \hat{\mu}(0) = \hat{\mathcal{U}_m}(0) = 1/m) \\
&\leq m^{3/2} \max_{i \neq 0} \{|\hat{\mu}(i)|\}.
\end{aligned}$$

Thus, for some $j \neq 0$, we have

$$|\hat{\mu}(j)| \geq \frac{2\epsilon}{m^{3/2}}.$$

Observe that

$$\begin{aligned}
\hat{\mu}(j) &= \frac{1}{m} \sum_{x \in \mathbb{Z}_m} \mu(x) \exp(-2\pi i j x/m) \\
&= \frac{1}{m} \mathbb{E}_{x \sim \mu}[\exp(-2\pi i j x/m)] \\
&= \frac{1}{m} \mathbb{E}_{x \sim \mu}\left[\left(\omega^{m-j}\right)^x\right].
\end{aligned}$$

Let $j' = m - j$. Thus, $|\hat{\mu}(j)| \geq \frac{2\epsilon}{m^{3/2}}$ implies that

$$\left|\mathbb{E}_{x \sim \mu}\left[\left(\omega^{j'}\right)^x\right]\right| \geq \frac{2\epsilon}{\sqrt{m}}.$$

This concludes the proof. □

B.2 Proof of Lemma 2.14

$$
\begin{aligned}
\epsilon^2 &\leq \left|\left(\sum_{w_1,w_2\in\mathbb{Z}_m^n} \mu_1\left(w_1\right)\mu_2\left(w_2\right)\left[\omega^{\langle w_1,w_2\rangle}\right]\right)\right|^2 \\
&\leq \left(\sum_{w_1\in\mathbb{Z}_m^n} \mu_1\left(w_1\right)\left|\sum_{w_2\in\mathbb{Z}_m^n} \mu_2\left(w_2\right)\left[\omega^{\langle w_1,w_2\rangle}\right]\right|\right)^2 \\
&\leq \left(\sum_{w_1\in\mathbb{Z}_m^n} \mu_1\left(w_1\right)^2\right)\left(\sum_{w_1\in\mathbb{Z}_m^n}\left|\left(\sum_{w_2\in\mathbb{Z}_m^n} \mu_2\left(w_2\right)\left[\omega^{\langle w_1,w_2\rangle}\right]\right)\right|^2\right) \\
&= \left(\mathrm{cp}\left(\mu_1\right)\right)\left(\sum_{w_1\in\mathbb{Z}_m^n}\sum_{w_2',w_2''\in\mathbb{Z}_m^n} \mu_2\left(w_2'\right)\mu_2\left(w_2''\right)\left[\omega^{\langle w_1,w_2'-w_2''\rangle}\right]\right) \\
&= \left(\mathrm{cp}\left(\mu_1\right)\right)\left(\sum_{w\in\mathbb{Z}_m^n} \mu_2\left(w\right)^2 m^n\right) \\
&= m^n\cdot\mathrm{cp}\left(\mu_1\right)\mathrm{cp}\left(\mu_2\right)
\end{aligned}
$$

□